

Department of One Health, Bioethics, and Technological Research

E-book

Ethical challenges in the use of artificial intelligence (AI) in medicine: human and non-human caring

Editors
Domenico Palombo
Rosagemma Ciliberti

DEPARTMENT OF ONE HEALTH, BIOETHICS, AND TECHNOLOGICAL RESEARCH

Prof. Domenico Palombo

Head of the Department of One Health, Bioethics, and Technological Research

STEERING COMMITTEE

Members of the Steering Committee:

- Professor Rosagemma Ciliberti (Genoa) Bioetichs (University of Genoa – Italy – ciliberti@unige.it
- Professor Lazar Davidovic (Belgrade) Dean of the Faculty of Medicine, Surgeon (University of Belgrade - Serbia) - davidovic.lazar@gmail.com
- Professor Ergun Demirsoy (Istanbul) Surgeon (University of Istanbul - Turkey) - ergundemirsoy@hotmail.com
- Manuel Dominguez Gonzales, Full Professor of Chemical Engineering at the University of Vigo (Spain): jmanuel@uvigo.es
- Dario Dongo Lawyer, President of Wiise srl (Italy): dario.dongo@me.com
- Professor Juan Carlos Parodi (Buenos Aires) Surgeon (University of Buenos Aires - Argentina) - parodijc@yahoo.com
- Patrizia Perego, Full Professor of Chemical Plants at the University of Genoa (Italy): p.perego@unige.it
- Ricardo Pinheiro de Souza Oliveira, Associate Professor at the University of São Paulo (Brasil): rpsolive@usp.br
- Professor Matthias Thielmann (Essen) Surgeon (University of Essen - Germany) - Matthias.Thielmann@uk-essen.de
-

ADVISORY COMMITTEE

Rosagemma Ciliberti (Coordinator)
ciliberti@unige.it

Linda Alfano
alfanolinda65@gmail.com

Alessandro Bonsignore
alessandro.bonsignore@unige.it

Antonio Rinaudo
antoniorinaudo@icloud.com

Valeria Schiavone
valeria.schiavone@gmail.com

SUMMARY

This index is constantly updated based on the acquisition of new contributions

Editors note (Domenico Palombo, Rosagemma Ciliberti)

Introduction
Artificial Intelligence: A Bridge to the New Future (Rui Nunes)

1. The role of artificial intelligences in medicine: opportunities, limits, and risks
§ Historical and critical briefs on the history of artificial intelligence (Valeria Schiavone)
§ The impact of AI on health care (Ergun Demirsoy)
§ What Artificial intelligences are and their main applications in medicine*

2. The "uncertain" knowledge of medicine
§ From evidence- based medicine to precision medicine: the critical and comparative approach of the physician *
§ Algorethics (Everardo Belloni)

Case report

3. Ethical principles for AI in healthcare
Communication with the patient using AI - informed consent (Chinmay Shah)
§The communication-driven care relationship under the impact of digital technologies and artificial intelligence developments (Patrizia Borsellino)
§ Ethics and Artificial Intelligence in the Doctor-Patient Relationship (Rosagemma Ciliberti; Linda Alfano).
§ Moral acting in AI*
§ AI and Mental Health: new challenges from the Ecotechnobioethics (Moty Benyakar; Nicolas Obiglio).

Cases report

4. Safety and public interest
§ The trustworthiness of AI and validation. Intelligibility and transparency *
§ AI management and responsibility's issues*
§ Data storage, sharing and governance *

Case report

5. AI and Surgery
§ The help of digital twins to enable customization of medical treatments *
§ AI assisted surgery*
§ The triadic relationship among the surgeon, the IA/Robot and the patient *
§ Computerized imaging management*

Case report

6. Education and Training

§ Ethical Frontiers in AI-Enabled Medical Education: Considerations on the Role of Large Language Models in Medical Learning (Gerson Hiroshi Yoshinari Júnior, Luciano Magalhães Vitorino)
§ Transforming Public Healthcare and Education Using AI-powered Mixed Reality Technology (Predrag Stevanović, Lazar Davidović)

Case report

7. The regulatory context

§ Digital, bioethics, deontology and law: which dialogue? (Sabina Semiz)
§ Governing Blueprint: Ethical AI in European Health Policy (Jasna Karačić Zanetti)
§Medico-Legal and Ethical Considerations about Artificial Intelligence in Healthcare: Brief Focus on the Italian Perspective (Alessandro Bonsignore, Francesca Buffelli)

Case report

8. The new challenges

§ Transhumanism's issues
§ Artificial Intelligence approaches to electrophysiological models of neurodegenerative disorders: technical aspects and ethical implications (Laura Carini, Sara Sommariva, Antonio Uccelli, Michele Piana)
§ Psychotherapy and artificial intelligence (Linda Alfano; Rosagemma Ciliberti)
§ Pros And Cons of OpenAI's ChatGPT (Chinmay Shah)
§ Embriología humana, o las matrices del CRISPR-Cas 9 en el campo de la genética*
§ For a new humanism in medicine

Case report

Bioethics Documents

- 1. CNB – CNBBSV Artificial intelligence and medicine: ethical aspects
- 2. Council of Europe. Mid-Term Review of DPPA Strategic Plan 2020-2022
- 3. Council of Europe. Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025)
- 4. International Telecommunication Union. Focus Group on Artificial Intelligence for Health (FG-AI4H) DEL01. Ethics and governance of artificial intelligence for health
- 5. Council of Europe. Report commissioned by the Steering Committee for Human Rights in the field of Biomedicine and Health (CDBIO)
- 6. European Parliament Artificial intelligence act
- 7. European Parliament Artificial intelligence in healthcare
- 8. WHO Regulatory considerations on artificial intelligence for health

Glossary

*** waiting for the acquisition of a contribution**

Editors note

The term "Artificial Intelligence" is both captivating and ambitious, marking a rapidly advancing and irreversible presence within healthcare: the realm of Digital Health.

Digital Health today encompasses various aspects ranging from Telemedicine to Cybersecurity, the use of Creative Intelligence to Digital Therapeutics, and Computational Labs where Big Data, through the application of Deep Learning, enables the creation of both clinical and managerial algorithms.

The bioethical implications of this Digital Revolution are becoming increasingly apparent, underscoring the vital importance of adopting a "remember to stay human" approach.
As it has been suggested, transitioning from algorithms to "androrithms" and incorporating the "human inside" is becoming crucial: from Algocracy to Algorethics!

It is important to note that this book unfolds a collection of contributions from experts across diverse fields, fostering an open and multidisciplinary dialogue. Furthermore, this work is designed to be a living document, consistently updated and expanded to reflect the ongoing developments in the field.

The introduction of AI solicits multiple ethical questions related, for example, to the new role and skills that physicians and healthcare professionals are called to assume, new forms of professional responsibility, training in medical schools and, finally, the dissemination and education of the population in the use of robotics and the most sophisticated technology.

The overall goal of this text is to serve as a valuable guide not only for healthcare professionals but also for members of various Professional Councils, including physicians, nurses, lawyers, philosophers, psychologists, and researchers. By offering insights from various perspectives, it aims to be a comprehensive resource that navigates the ethical landscape of AI in medicine and other professions, providing practical guidance and fostering a nuanced understanding of the complex challenges to be faced.

Domenico Palombo and Rosagemma Ciliberti

Artificial Intelligence: A Bridge to the New Future

Rui Nunes
Head of the International Chair in Bioethics

Artificial intelligence (AI) has dramatically changed the lives of people in the global space in which we move. Without realizing it, AI is already influencing important aspects of our social lives in areas such as the economy, the financial system, the creative arts, education, and even public health and healthcare delivery.

Also, scientific research and technological development are even today controlled on a large scale by artificial intelligence allowing for new patterns of innovation, such as proteomics, but at the same time putting serious ethical challenges related for example with generative artificial intelligence of which Chat GPT is a good example.

Therefore, I believe that artificial intelligence is not just another technological evolution such as the internet, or even an instrument that helps humanity in its search for economic and social development. But it is a true paradigm shift.

Why do I make this claim? First of all, it is true that, for the time being, artificial intelligence consists of software and hardware systems that act in the physical or digital dimensions. However, it has distinctive features that make it unique. Firstly, it has the capacity for learning and self-learning. Indeed, machine learning, deep learning, and the capacity to make associations between different concepts are the genetic fingerprints of this new era of AI.

Secondly, networking is another special characteristic of AI. It means that the autonomous activity of humanoid robots, for instance, is independent but also interconnected. The concept of "social robot" refers not only to interaction with humans but interconnectivity between different AI systems.

Also, AI robots can move in human physical space. So, AI is not enshrined in a computer anymore (hence the evolution from computational intelligence to AI) but can move and even contact humans by mastering human language. This is a great challenge to humanity because, as Yuval Harari rightly states, this enhanced capacity might capture the essentials of human culture and civilization and can capture the

operating system of humanity even before singularity is reached. That is a specific moment in time when AI systems overrule humans in the control of our common destiny.

Based on software that actuates in the virtual world (voice assistant, software for image analysis, search engines, facial recognition system, etc.) or incorporated in hardware devices, for example, advanced robots, autonomous vehicles, drones, etc. AI has apparently no technological limits.

Also, the challenge of AI systems in health both in healthcare delivery and in public health is paramount. It may be of use in promoting new treatment modalities besides preventing life-threatening diseases. Indeed, it may provide clinicians with a more accurate and detailed analysis by helping with the diagnosis and treatment of many diseases. Further, it may help assist caregivers in support of the elderly. It may be extremely useful for real-time monitoring of patients, sometimes at long distances. Telemedicine is a good example of this evolution. AI also has the potential to be of use in precision medicine and personalized healthcare.

The enormous potential of AI and its associated risks entail caution in its use. According to the Independent High-Level Expert Group on Artificial Intelligence (2019), all societies should use AI based on several guidelines:

1. Develop, deploy, and use AI systems in a way that adheres to the ethical principles of respect for human autonomy, prevention of harm, fairness, and explicability.
2. Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities, and others that have historically been disadvantaged or are at risk of exclusion, and to situations that are characterized by asymmetries of power or information, such as between employers and workers or between businesses and consumers.
3. Acknowledge that while bringing substantial benefits to individuals and society, AI systems may also pose certain risks and have a negative impact, including impacts that may be difficult to anticipate, identify, and/or measure (such as democracy, the rule of law, and distributive justice or on the human mind itself).
4. Ensure that the development, deployment, and use of AI systems meet the seven key requirements for trustworthy AI: (a) human agency and oversight; (b) technical robustness and safety; (c) privacy and data governance; (d) transparency; (e) diversity, nondiscrimination, and fairness; (f) environmental and societal well-being; and (g) accountability.

5. Foster research and innovation to help assess AI systems and to further the achievement of the requirements, disseminate results, open questions to the wider public, and systematically train a new generation of experts in AI ethics.
6. Involve stakeholders throughout the AI system life cycle. Foster training and education so that all stakeholders are aware of and trained in trustworthy AI. A fair and accountable use of AI in global health therefore implies robust ethical data governance.

These are some of the reasons why it is so important to promote a dispassionate ethical debate on artificial intelligence. This is why the book promoted by Prof. Domenico Palombo is so useful to all of us.

Indeed, the e-book *Ethical Challenges in the Use of Artificial Intelligence (AI) in Medicine: Human and Non-human Caring* addresses some of the most important issues in contemporary AI. Noticeably, this book is divided into different chapters namely the role of artificial intelligence in medicine: opportunities, limits, and risks, medicine: exact science, ethical principles for AI in healthcare, safety and public interest, AI and surgery, the psychosocial dimension, education and training, the regulatory context, and finally the new challenges of AI, such as the human body and the artificial body, transhumanism issues, chatbots and psychological support or even psychotherapy and artificial intelligence.

All these subjects are approached by a group of very differentiated scholars that through intellectual reflection as well as case reports translate to the medical field these complex issues.

It will be especially relevant for the International Chair in Bioethics to promote this important book in the international community so that humans are always in the loop of controlling their common destiny.

1. The role of artificial intelligences in medicine: opportunities, limits, and risks

Historical and critical briefs on the history of artificial intelligence

Valeria Schiavone
Experice Laboratory, Paris 8, University, France

The intention of this short text is not to retrace the history of artificial intelligence (AI), on which there is now an abundant bibliography, but to question its historical evolution, in some of its fundamental passages, from an ethical and bioethical point of view. Even before the famous ten Macy lectures (1), coordinated by Warren Mc Culloch, a founding event on the subject of brain inhibition took place in 1942 organised by Frank Fremont-Smith, administrator of the Josiah Macy Junior Foundation. The multi-disciplinary membership of the guests shows how the dawn of artificial intelligence corresponded with the interest of the entire scientific and cultural world in the functioning of human intelligence.
Participants were Warren McCulloch and Arthur Rosenblueth, neurobiologists and physicists, who first modelled the functioning of the neuronal cell and, later, of neuronal networks, according to binary logic. From their approach will develop the computationalist logic, according to which, AI will try to imitate the brain function of storing and logically-mathematical processing of information, resulting in the ability to compute and solve complex questions.

Margaret Mead and Gregory Bateson made an anthropological and psychological contribution, defending a systemic approach to the question of intelligence, pointing out its analogical, multimodal and complex functioning. We find precursor signs here, especially in the importance given to the interactions between systems, to the concept of embodied cognition (embodied mind) that contemporary research tends to integrate as harmoniously as possible with the computationalist model.

Finally, Frank Fremont-Smith and Lawrence Kelso Frank, administrators of the Josiah Macy Jr. Foundation, were present, partisans of the need to promote interdisciplinarity and mutual integration between the exact sciences and humanities in this field.
This very first conference also gave rise to a series of informal discussions in which Milton Erickson participated about hypnosis, and Howard Liddell about the conditioned reflex

The bioethical question is therefore already present here and, we might say, constitutes a sort of implicit red thread to the various disciplinary focuses that we could identify in this fundamental question: is it possible to find in the functioning of human intelligence, in its inseparable biological, cognitive and psychic aspects, an intrinsic purpose not only of preservation but of promotion of the living in all its forms? If so, artificial intelligence, which has always sought to imitate and in some way 'improve' human intelligence, would be faced with the challenge of reflecting in its functioning a kind of teleology, an intrinsic finality, which research into human intelligence and the biology of the cognitive processes of every living being would potentially unveil.
I therefore propose to follow a possible trace of the soundness of this hypothesis in the fundamental passages of the history of artificial intelligence, which, starting as early as Macy's lectures, shows a split in thought and ethical orientation between two currents within the scientific community invited to participate. We find, in fact, on the one hand a current that seeks only to reproduce cognitive processes through digital technology and, on the other, research that aspires to understand these processes in order to also highlight their psychological and social aspects. In any case, three major questions emerge from

Macy's lectures (1), which are investigated in different ways depending on the influence of the two different approaches: the first seeks to understand how perception and sensoriality can be simulated, summarising, at least initially, the recognition and processing of images and sounds; the second aims to identify, within the theory of games and systems, the criteria of unbeatability; and the third investigates, through symbolic calculation, all the possible applications of mathematical theorems.
These were already the areas outlined in July 1956 at the long eight-week working meeting at Dartmouth in the USA. There were twenty participants, including the four leading representatives of the Macy conferences: neurophysiologist Warren McCulloch, Julian Bigelow, pioneer of computer engineering, Claude Shannon, engineer and mathematician, and the British psychiatrist and pioneer of cybernetics, Ross Ashby. It was in July 1956 that John McCarty chose the term 'artificial intelligence' to distinguish it from the cybernetics of Norbert Wiener and John von Newman.
We are witnessing here another split that is both geographical and semantic and that is also worth analysing in its ethical implications. Artificial intelligence will begin to designate, especially in the United States, the search for the simulation of cognitive processes by means of complex computing machines, such as Alan Turing's. On the contrary, the cybernetics of the time will remain faithful to a European matrix and to the theoretical link with the theory of systems and the concept of homeostasis, already proposed by Claude Bernard, physician, physiologist and epistemologist, in 1865, in his Introduction to the Study of Experimental Medicine (2) and taken up by the biologist Ludwig von Bertanlanffy in 1968 in his General Theory of Systems (3).

Interestingly, McCarty, while remaining within a strictly computationalist model with respect to the functioning of human intelligence, emphasises the immediate philosophical scope of the issues related to the use of artificial intelligence. In the article, published together with Patrick J. Hayes, *Some philosophical problems from the standpoint of artificial intelligence* (4) he makes it clear that artificial intelligence needs philosophy to determine what can be defined as knowledge and what its characteristics are. Philosophy is thus fundamental to defining intelligence where bio-cognitive and psychosocial aspects alone merely trace a phenomenology of it. Philosophical reflection thus entails the natural attribution of value and purpose intrinsic to the living, human and non-human, which invests its environment intelligently, in a reciprocal adaptation and discovery of ever new strategies of 'being-in-the-world'. I refer here to the famous Heideggerian definition of human existence ad *'In-der-welt-sein'* (5), which today, following the suggestions of a holistic approach that bioethics makes its own, we can consider as relating to all living beings, bound together by sharing a common and interconnected horizon of life and meaning.

It is in the 1960s that we see a clear division of different and complementary aspects in the great field of the study of intelligence.
A Cybernetics, characterised by the robotics and automatism studies of Wiener and Newman, which is based on the retroactive response of the living (feedback) in its adaptation to the environment. In the field of cybernetics, we must mention the contribution of Francisco Varela (and his teacher Humberto Maturana), a Chilean neurobiologist, who revolutionised cognitive science with the concept of autopoiesis and enaction (6). In very synthetic terms, both these concepts emphasise that all complex systems (living organisms and social groups) have an autonomous generative capacity and that the body, far from being a mere container of the mind, rather produces it as a kind of emergence of consciousness from cognitive-neural processes.
B Ludwig von Bertalanffy's general systems theory (inspired, as we have seen, by the work of Ross Ashby), which is based on two fundamental principles: the recognition of a close interconnection and mutual significance between each element and the whole; and the idea that 'the whole is not the mere sum of the parts', but a biological, cognitive and social entity by itself. Also referable to this theory of

complex systems is the idea of complex thinking inaugurated by Ross Ashby and taken up by Edgar Morin in relation to the unity of the human being in the famous Royaumont Colloquium of 1972.

C Artificial intelligence as engineering, based on the analogy between the brain's functions and the digital, computational capabilities, which develops increasingly precise and complex machines, languages and ways of representing data and processing them.

D Classification methods: their purpose is either to define, in the best possible way, a group of objects from an already defined language; or to find languages and modes of representation that are more and more suitable for optimal classification. Classification methods are based on descriptive statistics following Jean-Pierre Benzécri's works (7), a French mathematician and statistician specialising in data analysis; and on single- and multi-layer neural and multi-layer neural networks, referring to Yan LeCun's work (8) on deep learning and on Bayesian inference statistical tools (a method of inference by which the probabilities of various hypothetical causes are calculated from observation).

This brief and non-exhaustive historical excursus should have shown how, from its origins, the study of artificial intelligence was intrinsically linked to philosophical and ethical questions related to the common inhabitation of the world by all living beings. Especially in the field of medicine, the progress that artificial intelligence has made and continues to make in the four main areas identified above is evident. It is therefore crucial to consider, in each of these areas, all the ethical and philosophical implications: from complexity thinking, to the processes of emergence linked to the embodied mind, to the question of the best classification and treatment of data.
All these issues, even in their early developments, have been seen as strictly linked to the fundamental questions of the interconnections between human beings, the environment and other living beings, and the need to think in a complex way in order to guarantee, each time, an ethical application of artificial intelligence in medical science.

Therefore, in my opinion, it is not a question of worrying about the progress of artificial intelligence, but of continuing to ask the fundamental questions:

- how to ensure that artificial intelligence continues to mimic human intelligence in all its biological, cognitive and psychosocial characteristics.
- how to orient all further developments of artificial intelligence towards the ethics that complex thinking suggests to us, so that its applications are not reductive of human richness and its intrinsic connections with the richness of all living beings.

References

1. Klaus Pias (edited by), *Cybernetics, The Macy Conferences 1946-1953. The Complete Transactions,* Diaphanes 2016
2. Claude Bernard, *Introduction à l'étude de la médecine expérimentale,*1865. (OCLC 600479635) (Rééd. Champs, Flammarion, (ISBN 2080811371)).
3. Ludwig von Bertanlanffy, *General System theory: Foundations, Development, Applications,* New York: George Braziller, revised edition 1976: (ISBN 0-8076-0453-4)
4. John McCarty, *Some philosophical problems from the standpoint of artificial intelligence, in* Meltzer, B., and Michie, D., eds., *Machine Intelligence* 4, Edinburgh University Press. 463-502, 1969.
5. Martin Heidegger: *Sein und Zeit.* 11. Auflage. Niemeyer, Tübingen 1967
6. Pierre de Loor, Kristen Manac'H, Jacques Tisseau. *Intelligence Artificielle Basée sur l'Enaction: et si l'homme était dans la boucle?*. 2013, 10.1007/s11023-009-9165-3 hal-00654120

7. J-P Benzécri et al. *L'analyse des données*, Dunot 1973.Inaugural Yan LeCun's lecture at the Collège de France on deep learning UPL7915574462521283497_lecun_20160204_college_de_france_lecon_inaugurale.pdf (college-de-france.fr)

The impact of AI on health care

Professor Ergun Demirsoy Chief of Cardiac Vascular Surgery Department-Şişli International Kolan Hospital

The development of artificial intelligence (AI) has triggered a discussion about the changing roles of physicians and health professionals in the healthcare industry as it has started to transform the way healthcare is being delivered around the globe. It seems that it can revolutionize the process and improve patient outcomes by boosting productivity if handled with caution. Artificial intelligence cannot be defined as one technology, but rather a collection of them. Most of these technologies have immediate relevance to the healthcare field, but the specific processes and tasks they support cover a wide spectrum. In order to effectively integrate AI into the healthcare industry, it is necessary for physicians and health professionals to adapt to these changing roles and embrace them. Additionally, healthcare organizations will need to invest in training and education programs to help physicians and other health professionals develop these skills.

One of the primary ways in which AI is changing the roles of healthcare professionals is by increasing access to medical information. Healthcare data is often fragmented and is in various formats. By using AI and machine learning technologies, organizations can connect distinct data to get a more accurate picture of the individuals behind the data. Additionally, the internet has already made it possible for patients to research health conditions and treatments, and AI is taking this one step further by analyzing medical data to help clinicians make more informed decisions. This means that physicians and other health professionals will need to become more capable of interpreting data and using AI-based tools to make diagnoses and develop treatment plans.

AI is also changing the traditional roles of physicians and other health professionals by providing new opportunities for collaboration. For example, AI-powered tools can help physicians communicate more effectively with patients and provide them with information and advice. By encoding clinical guidelines or existing clinical protocols through a digital system often provides a baseline , which then can be broadened by models that learn from data. We see that more data has started to talk to each other which in turn is helping the healthcare professionals to make informed decisions. At this point, the ultimate goal is the gathering of connected data. Additionally, AI can enable remote monitoring of patients, allowing healthcare professionals to assess patient conditions without being physically present.

Another way in which AI is changing the roles of healthcare professionals is by automating routine tasks. AI can automatically scan electronic health records to identify medications, dosages, and allergies. Similarly, AI can automate many aspects of medical billing and coding, freeing up physicians and other health professionals to focus on providing quality care to their patients.

While the rise of AI in healthcare presents numerous opportunities, it also raises concerns about the impact on the roles of healthcare professionals. Many are worried that AI and other technology could replace human healthcare professionals, leading to job losses and a decline in the quality of care. For example, if AI-enabled diagnostic tools become more prevalent, the need for human diagnosticians may decrease. Physicians and healthcare professionals must work together to ensure that the introduction of AI into healthcare does not create undesirable impacts on employment. However, others argue that AI can never fully replace the human touch when it comes to delivering healthca re, and that physicians and other health professionals will always be needed to provide emotional support, comfort, and personalized

care to their patients. As well as the benefits, there are also some challenges to adopting AI in healthcare, including having to meet regulatory requirements and overcoming trust issues with machine learning results. With the increasing use of AI in healthcare, vast amounts of patient data are being collected and analyzed. While this data can be beneficial for patient care, there are concerns around the potential for breaches of patient privacy. Physicians and healthcare professionals must ensure that patient data is collected, analyzed, and stored in a secure and ethical manner.

Additionally, there are concerns around potential biases in AI algorithms used in healthcare. AI algorithms are only as good as the data that is inserted into them. If the data used to train these algorithms is biased, the algorithms themselves will be biased. This can lead to potential errors or misdiagnoses, and can ultimately have negative impacts on patient care. Physicians and health professionals must pay close attention to the data and algorithms they are using and ensure that they are objective, accurate, and unbiased.

Another challenge facing physicians and health professionals in the context of AI is ensuring that AI is augmenting human expertise, rather than replacing it. While AI can be incredibly helpful in automating routine tasks and providing physicians with access to large amounts of data, it cannot replace the human element of patient care. Patients still need the compassion, connection, and expertise that only a human physician or healthcare professional can provide. Physicians and health professionals must ensure that AI is used in a way that supports and enhances human expertise, rather than replaces it.

In conclusion, the changing roles of physicians and health professionals in the context of AI present significant challenges as well as numerous exciting opportunities for improving patient care. It also seems clear that AI systems will not replace human clinicians on a large scale, but rather will augment their efforts to care for patients. Prevention has become increasingly important over the last several decades and healthcare providers and institutions have been quite successful in achieving desired outcomes as a result of an increase in focus on disease prevention which in turn helped to manage costly medical conditions as well. Therefore, the combination of improved quality of life for patients and system efficiency has brought physicians and policy experts together with the mutual goal of building a robust healthcare prevention infrastructure.While adapting to new technologies, ensuring the ethical use of patient data, addressing algorithmic bias, and ensuring that AI enhances rather than replaces human expertise are all important challenges, when approached correctly, AI has the potential to revolutionize healthcare delivery and improve patient outcomes. Physicians and healthcare professionals must embrace these changes and work together to ensure that AI is used in an ethical, responsible, and effective manner.

LIST OF REFERENCES

1. Jha, S., Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. JAMA, 316(22), 2353-2354.

2. Powles, J., Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. Health and Technology, 7(4), 351-367.

3. Wachter, R. (2018). The digital doctor: Hope, hype, and harm at the dawn of medicine's computer age. McGraw Hill Professional.

4. Obermeyer, Z., Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.

5. Lyu, Y., Zhang, P., & Liu, M. (2019). The impact of artificial intelligence on healthcare: From diagnosis to treatment. IEEE Intelligent Systems, 34(5), 49-54.

6. Kulendran, M., Lim, M., Laws, G., Chow, A., Nehme, J., Darzi, A., & Purkayastha, S. (2019). Artificial intelligence in healthcare: past, present and future. British Journal of Surgery, 106(11), 1521-1528.

7. Kamaleswaran, R., Chan, A., & Trachtenberg, J. (2020). Artificial intelligence in healthcare: current perspectives and future directions. American Journal of Medicine, 133(2), 221-227.

8. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

9. Rimmer, A. (2020). Artificial intelligence is changing the role of doctors, but it will not replace them. BMJ, 368, m340.

10. Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. Journal of the American Medical Association, 318(6), 517-518.

2. The "uncertain" knowledge of medicine

Algorethics

Everardo Belloni
Adjunct professor at POLIMI Graduate School of Management

1. Algorethics main concept

The concept of Algorethics represents a new challenge in ethical reflection that wants to outline some ethical principles to be translated or implemented in software in a view to mitigate the unintended effects of algorithmic execution.

Algorethics can be defined as a **set of guidelines** that advise on the design and outcomes of artificial intelligence (AI). As instances of unfair AI outcomes have come to light, new guidelines have emerged, primarily from data science communities, to address concerns around the ethics of AI. As the appropriate expertise develops within the government industry, we can expect more AI protocols for companies to follow, enabling them to avoid any infringements on human rights and civil liberties.

2. Concerns with AI algorithms

Trust is essential in the value exchange that occurs between a consumer/citizen and an organization. Consumers who do not trust how their data is being used and worry that it is being used to disadvantage them in some way, are at risk of withdrawing from the relationship with a brand.

The following stats illustrate that trust in the ethical use of AI is incredibly important to consumers and they are willing to reward brands they view as such with their continued business, according to a Capgemini Research Institute report (2019): 62% would place higher trust in a company whose AI interactions they perceived as ethical; 59% would have higher loyalty to the company; 55% would purchase more products, providing positive feedback on social media; 61% would share positive experiences with their peers.

A number of issues surrounding AI technologies must however be addressed:

- **Technological singularity or superintelligence.** This concerns the possibility of any generative AI[1] system or algorithm that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills.
- **AI impact on jobs.** With every disruptive, new technology, we see that the market demand for specific job roles shift. AI should be viewed in a similar manner, causing to shift the demand of jobs to other areas. On a greater scale, this process can threaten the overall organization of the human society, generating the need for new management forms to help managing the AI potential negative impacts and the resulting complex problems as data grows and changes every day.
- **Privacy.** Privacy tends to be discussed in the context of data privacy, data protection and data security, and these concerns have allowed policymakers to make more strides here in recent years[2].

[1] The **generative AI** is a type of artificial intelligence that can interact with users in natural language and create novel data and contents, ranging from story outlines, reports, and other text outputs to multimodal content like images, videos, and audio.

[2] For example, in 2016, GDPR legislation was created to protect the personal data of people in the European Union and European Economic Area, giving individuals more control of their data: in particular Article 22 of the regulation includes a 'right to explanation', so-called because organizations must be able to provide 'meaningful information about the logic involved' in automated decisions. Canada has published a Directive on Automated Decision-Making. The Directive, a key pillar of the country's commitment to ethical AI practices, centers around the Algorithmic Impact Assessment (AIA), a tool that determines exactly what kind of human intervention, peer review, monitoring, and

This recent legislation has forced companies to rethink how they store and use personally identifiable data. As a result, investments within security have become an increasing priority for businesses as they seek to eliminate any vulnerabilities and opportunities for surveillance, hacking, and cyberattacks.

- **Bias and discrimination.** Bias and discrimination to which are prone intelligent systems have raised many ethical questions regarding their use. How can we safeguard against bias and discrimination when the training data, upon which AI systems are built, are inevitably biased? Moreover, bias and discrimination can be found in a vast number of applications from facial recognition software to social media algorithms.
- **Accountability.** At present, there is not significant legislation to regulate AI practices, posing real enforcement mechanism to ensure that ethical AI is practiced. There is evidence showing that the combination of distributed responsibility and lack of regulatory oversight and societal controls into potential consequences isn't necessarily conducive to preventing harm to society.

3. Main principles for AI ethics

While rules and protocols develop to manage the use of AI, the academic community has leveraged the Belmont Report[3] as a means to guide ethics within algorithmic development. Three principles of the Belmont Report have served as a foundation guide for experiment and algorithm design:

- **Respect for Persons**: Individuals should be aware of the potential risks and benefits of any experiment that they're a part of, and they should be able to choose to participate or withdraw at any time before and during the experiment.
- **Beneficence**: Despite the intention to do good, designers should commit themselves not to cause harm in developing artificial intelligence where algorithms can amplify biases around race, gender, political leanings, et cetera.
- **Justice**: This requires distributing burdens and benefits with fairness and equality, following five guiding principles: equal share, individual need, individual effort, societal contribution and merit.

Both in theory and in practice, there are several questions to consider when evaluating whether an AI solution is ethical:

- **Does the solution deliver fair and equitable outcomes?** The ultimate objective in delivering machine learning and AI solutions should be to avoid building systems that create or reinforce inequalities among humans.
- **Does the solution introduce or exacerbate bias?** Bias is often an unfortunate fact of life and is undesirable when it increases inequality or unfairly favors one group over another. However, bias may be acceptable if knowingly applied to rectify larger environmental social distortions, but if left unnoticed, bias can become systematically amplified or reinforced. There are three levels in which

contingency planning any AI tools designed to serve citizens will need. The US Federal Government has also addressed the issue in proposed legislation around automated-decision systems. Similarly, the Australian Human Rights Commission is conducting wide consultations on AI-informed decision making. In the United States, individual states are developing policies, such as the California Consumer Privacy Act (CCPA), which require businesses to inform consumers about the collection of their data.

[3] The Belmont Report was written by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Commission, created as a result of the National Research Act of 1974, was charged with identifying the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and developing guidelines to assure that such research is conducted in accordance with those principles. Informed by monthly discussions that spanned nearly four years and an intensive four days of deliberation in 1976, the Commission published the Belmont Report, which identifies basic ethical principles and guidelines that address ethical issues arising from the conduct of research with human subjects.

bias can occur in AI systems: at the data level regarding the way data is collected (sampled or selected); at the algorithm development level; and at the deployment level. Across these three levels, AI's tendency to scale up embedded prejudice can be worrying by turning mere correlation insights into causative outcomes.

- **Will this solution result in humans feeling or experiencing loss of control or agency?** This reflects the fear that, as AI-driven systems become more pervasive, humans will lose their ability to decide for themselves (evaluate alternative options or have the freedom to act), since having the ability to direct their own actions without undue influence is something all humans expect to be able to do. To mitigate these concerns, AI developers should consider how humans will interact with each AI solution and define the engagement accordingly, by clearly communicating when AI systems are in place, how they are used and, if appropriate, allowing people to opt-out, intercede, customize or challenge algorithmic actions or decisions
- **What is the impact on existing roles and employees?** The impact of AI on existing roles and the need to modify established business practices must be adequately addressed. Existing resources will need to be upskilled or redeployed, though some other jobs may be eliminated. Beyond changing the nature of existing work, AI will also require employees to become more technically literate.

4. How to establish an AI ethics framework

Researchers have started to assemble **frameworks and concept**s to address some of the current ethical concerns and shape the future of work within the field. Overall, there is some consensus around incorporating the following elements:

- **Governance.** Companies should leverage their existing organizational structure to help manage ethical AI, by including ethical principles to their data collection established governance systems, the aim of which is mainly to facilitate data standardization and quality assurance.
- **Algorithms Perspicuity.** Machine learning and deep learning models are frequently "black box models" as it's usually unclear how they allow at a given decision. Transparency should seek to eliminate this ambiguity around model assembly and model outputs by allowing for a human understandable explanation that expresses the rationale behind an algorithm. If we can better understand the why, we will be better equipped to avoid AI risks, such as bias and discrimination.

Emerging governance practices include **management review boards** to vet proposed applications, implementing model development standards that incorporate frequent checkpoints with diverse stakeholders, routine monitoring and review of results and outcomes, communicating where AI is being deployed and providing recourse for those impacted to understand and/or appeal decisions made by automated systems.

In detail, specific items of an **ethically aligned design** of AI system are as follows:

- **Human Rights**: AI solutions shall be created and operated to respect, promote, and protect internationally recognized human rights.
- **Well-Being:** AI creators shall adopt increased human well-being as a primary success criterion for development.
- **Transparency:** The basis of a particular AI decision should always be discoverable, according to a perspicuity requirement.
- **Accountability**: AI shall be created and operated to provide an unambiguous rationale for all decisions made.
- **Awareness of Misuse:** AI creators shall guard against all potential misuses and risks of AI in operation.

- **Competence:** AI creators and operators shall adhere to the knowledge and skill required for safe and effective operation.
- **Data Agency:** AI creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
- **Effectiveness:** AI creators and operators shall provide evidence of the effectiveness and fitness for purpose of the AI systems deployed.

5. Organizations on ethical AI and Algorethics

A number of organizations have emerged to promote ethical conduct in the field of artificial intelligence and machine learning algorithms. The following organizations and projects provide resources on implementing ethical AI:

- **The World Economic Forum** aims to bring together the public and private sectors to co-design, test, and implement policies that increase the benefits of artificial intelligence and machine learning, while developing projects to protect vulnerable people.
- The **OECD AI Principles** focus on how governments and other actors can shape a human-centric approach to trustworthy AI. As an OECD legal instrument, the principles represent a common aspiration for its adhering countries.
- **The EU AI Act** is a proposed European law on AI. The law assigns applications of AI to three risk categories: applications and systems that create an unacceptable risk will be banned; high-risk applications will be subject to specific legal requirements; applications not explicitly banned or listed as high-risk will be largely left unregulated and subject only to general ethical principles.
- **IEEE Transactions on Artificial Intelligence**: within its Technical Committee for Ethical, Legal, Social, Environmental and Human Dimensions of AI/CI (SHIELD), it focuses on proposing technical /practical solutions to assess the impact of AI across multiple intertwined dimensions such as ethical, economic and societal.
- **Algorithm Watch**: This is a non-profit research and advocacy organization that is committed to watching, unpack and analyze automated decision-making (ADM) systems and their impact on society.
- **AI Now Institute**: Founded in 2017, it produces diagnosis and policy research to address the concentration of power in the AI industry with reference to social implications of artificial intelligence.
- **CHAI**: The Center for Human-Compatible Artificial Intelligence is a cooperation of various institutes and universities to promote trustworthy AI and provable beneficial systems. CHAI's goal is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems, addressing the problem of control: given that the solutions developed by such systems are intrinsically unpredictable by humans, it may occur that some such solutions result in negative and perhaps irreversible outcomes for humans. CHAI's goal is to ensure that this eventuality cannot arise, by refocusing AI away from the capability to achieve arbitrary objectives and towards the ability to generate provably beneficial behavior. Because the meaning of beneficial depends on properties of humans, this task inevitably includes elements from the social sciences in addition to AI.
- **NASCAI**: The National Security Commission on Artificial Intelligence is an independent commission addressing methods and means necessary to advance the development of artificial intelligence, machine learning and associated technologies to comprehensively address the national security and defense needs of the United States.

- **Future of Life Institute's Asilomar AI Principles**: Its mission is to steer transformative technologies away from extreme, large-scale risks and towards benefiting life.

References

https://www.weforum.org/topics/artificial-intelligence-and-robotics
https://artificialintelligenceact.eu/
https://oecd.ai/en/ai-principles
https://cis.ieee.org/publications/ieee-transactions-on-artificial-intelligence
https://artificialintelligenceact.eu/
https://algorithmwatch.org/en/
https://www.nscai.gov/
https://ainowinstitute.org/
https://humancompatible.ai/
https://futureoflife.org/

3. Ethical principles for AI in healthcare

“Communication with the patient using A.I. - informed consent”

Dr. Chinmay Shah
Professor & Head, Department of Physiology, Government Medical College, Bhavnagar, Gujarat, India. cjshah79@yahoo.co.in, Orcid ID : 0000-0002-4714-0129

Abstract: Doctor patient communication pays a vital role in healthare delivary. As such communication in health care is considered amongts one of the diffircult conversaion. Additon of AI in communication may create trouble due to insufficient knowledge of AI on either side. Out of all communication, informed consent process has dual importance i.e ethical as well as by virtur of law, thus specific care must be taken in process of informed consent. In this chapter we are going to discuss regarding care need to be taken while using AI in communication with specific focus on informed consent process.

1. Introduction:

Nowadays Artificial intelligence is used in a big way in health care, starting from appointment of physician till discharge. There are both positive and negative side of increasing use of technology and particularly AI in health care.

Day to day communication is playing crucial role in patient doctor relationship. Assault on doctors and health care organization has warn us to learn art of communication which is appropriate for circumstances and satisfy emotional and psychological need to patient. When healthcare worker are still in learning phase for socio behavioral change communication in regards to healthcare, addition of AI may worsen the situation. Thus, we need to keep in mind to use AI for communication with Patient only whenever indicated.

Communication with patient is needed for

1. Booking of appointment
2. Taking history
3. To give instruction for examination
4. To order a battery of test
5. To answer patient question in regards to symptoms or investigation
6. To take informed consent for any procedure for treatment or for research
7. To give instructions during medication and surgery
8. During discharge
9. During follow up
10. Billing and Medical Insurance

Out of all scenario mentioned above booking of appointment, order a battery of test , billing and medical insurance are comparatively easy communication as involved minimal intervention, thus AI can be easily implemented for this purpose. All other component of patient communication require more or less dialog so it will be little difficult to use AI for that instance, but we have started using it for our convenience. Medical ethics has begun to highlight concerns about uses of AI and robotics in health care, including algorithmic bias, the opacity and lack of intelligibility of AI systems, patient-clinician relationships, potential dehumanization of health care, and erosion of physician skill[1].

2. Sensitizing stakeholder

Interconnected with lack of knowledge about AI systems amongst both patient and health care professionals may create error. We need to keep in mind psychology of patient as well as healthcare worker while implementing AI in healthcare.

Following stakeholder need to sensitize for effective use of AI in Patient Communication[2]:

- *Coders and designers*. They play vital role in creating AI software/platform / interphase. They must document what they created and also provide accurate user guide regarding implementation of same.
- *Medical device companies*. Further Companies should clearly articulate prerequisites for successful application of an AI technology, such as the quality of diagnostics, imaging, and preparation for surgical procedures.
- *Physicians and other health care professionals' communication*. Physicians should be responsible for acquiring basic understanding of the AI devices they use and the types and likelihood of errors across subgroups, insofar as this information is available. Physicians are responsible for communicating relevant information to patients and health care teams and adhere to standards provided by device companies.
- *Hospitals and health care systems*. Hospitals are key to ensuring proper development, implementation, and monitoring of protocols and best practices for use of AI systems in health care. This organizational responsibility includes providing training, protocols, and best practices related to AI use and properly informing patients about the technology. Hospitals should also be involved in developing robustness measures (including simultaneous diagnosis and crosschecking by physicians and AI).

3. AI and Informed Consent

Once all stakeholders are sensitized, it will be advisable to start using AI in Healthcare. Still out of all, one communication is extremely important that is informed consent, it is important by virtue of ethics as well as law. Thus we need to give specific attention on it. Let's see few important aspects of the same.

As mentioned previously we need to create point to point script for our coder and designer so that they do not miss information about a proposed procedure/test/ treatment, its benefits and risks; and any alternative options; subjects right and responsibility ; information regarding personal data protection and finally details regarding compensation and cost.

Further AI should provide opportunity to ask question to highly learned chatbot(AI). With this knowledge, the patient decides to either consent or not consent to the recommended plan.

Schiff[3] mentioned in his article that "for an informed consent process to proceed appropriately, it requires physicians to be sufficiently knowledgeable to explain to patients how an AI device works."[3]

Those who plan to use these technologies in practice should be able to[4]:

- Provide patients with a general explanation of how the AI program or system works
- Explain the healthcare provider's experience using the AI program or system

- Describe to patients the risks versus potential benefits of the AI technology (e.g., compared to human accuracy)
- Discuss with patients the human versus machine roles and responsibilities in diagnosis, treatment, and procedures
- Describe any safeguards that have been put in place, such as cross-checking results between clinicians and AI programs
- Explain issues related to confidentiality of patient's information and any data privacy risks

For an informed consent process to proceed appropriately and accurately, Physicians who use machine-learning systems can become more educated about their construction, the data sets they are built on, and their limitations, they also need to be sufficiently knowledgeable to explain the patients how an AI device works in informed consent process, as the presentation of information using AI can be made complicated by possible patient and physician fears, overconfidence, or confusion. Remaining ignorant about the construction of machine-learning systems or allowing them to be constructed as black boxes could lead to ethically problematic outcomes.[5] Health care organization must plan regular training of all health care provider for the same.

While using AI in taking consent organizations need to build trust amongst patient by being transparent about the purpose behind the use of the technology, What data is collected, processed, and used, Measures taken to safeguard the security and privacy of personally identifiable data and Any known issues, data breaches, safe practices to follow. They must Empower patients with Means to seek more information/explanation about the technology and data collection, Control over their data and Ability to seek recourse if things go wrong[6].

4. Reference:

1. Barbash GI, Glied SA. New technology and health care costs—the case of robot-assisted surgery. New Engl J Med. 2010;363(8):701-704.

2. Schiff, D., & Borenstein, J. (n.d.). CASE AND COMMENTARY how should clinicians communicate with patients about the roles of artificially intelligent team members? Ama-assn.org. Retrieved November 9, 2023, from https://journalofethics.ama-assn.org/sites/default/files/2019-01/cscm3-1902_0.pdf

3. Schiff, D., & Borenstein, J. (2019, February). How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA Journal of Ethics, 21(2), E119-197. doi: 10.1001/amajethics.2019.138

4. Artificial Intelligence and Informed Consent. (n.d.). Medpro.com. Retrieved November 9, 2023, from https://www.medpro.com/artificial-intelligence-informedconsent

5. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981-983.

6. Capgemini Research Institute, "Championing Data Protection and Privacy: a source of competitive advantage in the digital century," September 2019.

The communication-driven care relationship under the impact of digital technologies and artificial intelligence developments

By Patrizia Borsellino
Full Professor of Philosophy of Law and Bioethics
University of Milano-Bicocca – Milan, Italy
e-mail: patrizia.borsellino@unimib.it

Abstract:
After recalling the profound transformations that have affected medicine since the second half of the twentieth century, the chapter draws attention to the model of the therapeutic relationship based on the involvement of doctors and patients in decision-making, each for their part and with their rightful role, identifying communication as the necessary condition for the implementation of care appropriate to the patient's situation and respectful of his or her wishes. There is, however, a further transformation, by which medicine has already been invested and will increasingly be so in the years to come, to be placed at the centre of the reflection on the care relationship. This is linked to the progress made in the field of information technology and the development of digitalisation processes, thanks to which it has become possible to both collect and store an enormous amount of data, making them easily accessible, and to process them by means of (increasingly) intelligent computer programmes, the so-called machine learning algorithms. Subject of specific consideration in the essay is the question of the impact of the digitalisation of health data, and their algorithmic processing by means of artificial intelligence technologies, on the care relationship and clinical decision-making process and, in particular, on the communicative dimension that constitutes its backbone. In the light of the clarification of the notion of communication and its relationship with the notion of information, as well as the reasons that justify the close relationship between the care relationship and communication, the analysis shows that information, amplified and speeded up by the new technological resources, can only play a key role in transmitting knowledge, on which to base a good clinical decision, if it is brought back within the care relationship as a two-way communicative context, dialogic and non-monologic, as well as marked by empathy and trust. Only by operating within such a context will the physician be able to scale down the idea of self-sufficiency or, at the very least, of the necessary prevalence of the algorithm over further informational elements that have emerged in the confrontation, which are important in order to clearly envisage the therapeutic alternatives that can be proposed and will be able to support the patient in making an informed choice.

1. Introduction. The care relationship reshaped by communication against the background of medicine's increased possibilities of intervention in the second half of the twentieth century

In the course of the twentieth century and, above all, from its second half, following extraordinary scientific advances, medicine has undergone profound transformations. It has, in fact, taken shape as a dynamic context, characterised by the ability to successfully intervene on an ever increasing number of pathologies, and, in any case, to affect human life in the various phases of its unfolding, and, above all, to modulate the time and manner of its conclusion, thanks to the growing availability of methods and therapeutic strategies functional to the prolongation of survival.

In this scenario of increased, and at times problematic, possibilities of intervention, the problem of the criteria to be applied in making therapeutic choices, and the problem of who should rightly be considered to be responsible for these choices, came to the fore. This came about when there was widespread awareness of the need to critically rethink the traditional 'paternalistic-vitalistic' representation of the care

relationship, which, for a long time, took for granted both the attribution of decisions to physicians alone and their finalisation to the prolongation of survival, whatever the cost.
Thanks to the questioning of this model on the theoretical level[1], but also to its overcoming on the deontological and juridical level[2], the contours of the care relationship as a metaphorical place of confrontation (necessary) and encounter (desirable) between the subjects involved in the clinical decision-making process, each for the part and with the role that he or she is entitled to, have become increasingly clear. On the one hand, the physicians, invested, by virtue of their professional competence, with the prerogative of identifying and proposing the appropriate treatment pathways for patients' conditions, in a context of increased possibilities of intervention, but not of decreased margins of uncertainty. On the other hand, the recipients of the interventions, recognised, as a rule, as holders of 'decision-making autonomy'[3], i.e. the prerogative of having the last word on treatments, whether they are treatments to be implemented or treatments already implemented and intended to continue over time, by expressing their consent, in the case of adherence to the therapeutic proposal, or by expressing their disagreement, in the case of non-acceptance.
This is a model that enhances the irreplaceable role of the caregivers, but, at the same time, requires a reshaping of their 'modus operandi', which goes hand in hand with the recognition of the active role invested in the patient in the treatment decision-making process. In the context of the therapeutic relationship thus understood, the 'good' of the patient, destined to guide the conduct, is no longer, in fact, the one with pretended objective connotations, a priori identified with the prolongation, at all costs, of survival, but rather, the one declined in terms of 'adequate' response to the needs and expectations of the patient, to be identified by establishing, first of all, with the patient himself, and, if he wishes, also with the people close to him, a communication, which the doctor is required to consider an integral part of the treatment.
Now, the idea that there must be a close and significant relationship between the care relationship and communication has been, for some years now, increasingly acknowledged on a theoretical level[4], but, in some contexts, such as, for example, the Italian one, it has also received significant confirmation on a regulatory level[5], starting from the conviction that communication can play a key role in promoting the implementation of practices that conform to the above-mentioned model of the care relationship, and can have a positive impact on a wide range of critical issues that the care relationship must deal with.

2. New scenarios of transformation of medicine. What impact on the care relationship

Reflection on the care relationship and on the best model for it, however, must now take into account the further transformation that medicine has already undergone and will increasingly undergo in the years to come.
The transformation in question is that linked to the extraordinary and rapid progress made in the field of information technology and the development of digitalisation processes, thanks to which it has become possible both to collect and store, making them easily accessible, an enormous amount of data, and to process them using (increasingly) intelligent computer programmes, so-called “machine learning algorithms”, which are able to create predictive models, and to derive answers, and thus solutions for even new problems, from the analysis of the data, guaranteed by the breadth of the correlations on which they are based, and characterised by an increasing degree of autonomy from human (programming) intervention.
The above-mentioned technological developments have already profoundly affected the organisation of healthcare structures and services, have made possible the provision of services at a distance inherent in telemedicine, and have accompanied and enabled the extraordinary progress achieved in recent years in various areas of specialisation, from that of radiodiagnostics to oncology, fuelled by the contributions of immunology and genetics, to surgery enriched by robotics, to name but a few in which there have been

results and improvements of such relevance as to dispel the mistrust or aversion to the applications of these advanced frontiers of technology in the field of human health[6].
But what can be the impact of the digitisation of health data, and their algorithmic processing by artificial intelligence technologies, on the care relationship and clinical decision-making?
More specifically, can one consider that 'data-driven medicine'[7], and the entry into the clinical context of software offering the virtual simulation of diagnostic responses and intervention strategies, call into question the aforementioned model of the therapeutic relationship, in which communication constitutes the necessary condition for the implementation of care appropriate to the patient's situation and respectful of his or her wishes, or can one put forward arguments in support of the need to safeguard that model, and to identify the paths to be taken to make it compatible with the new (non-reversible) technological scenarios?

3. The reasons for communication in the care relationship

The answer to those questions requires the careful consideration of the characteristics and applicative potentialities of the new information and artificial intelligence technologies, but it also requires that we dwell, preliminarily, on the reasons that support the idea of a close and significant relationship to be established between the care relationship and communication, and, first of all, that we clarify the meaning in which the terms "communication" and "communicating" are used in relation to care situations.
Now, the meaning of "communication" on which it is possible and opportune to converge in a perspective attentive to the contributions of linguistics and the theory of language, as well as psychology, is that according to which communication does not consist in the unidirectional transmission of something from someone to someone else, but rather in a two-way process between an issuer who is, in turn, a receiver of messages, and a receiver who is, in turn, an issuer. A meaning that harks back to the notions of dialogue, exchange, reciprocity and, therefore, to the idea of sharing and putting something in common, which the Latin verb 'communicare' already drew from the expression 'cum munus', to perform a function/bear a load together, in which it had its etymological root.
Unlike communication, information is the transmission of data, carried out, with the aim of transferring knowledge, by means of a unidirectional process. While communication is a form of interpersonal relationship between actors who, in a dialogical relationship, and starting from their subjectivities, cooperate in defining the discursive contents emerging from the communication itself, information is (or should be) an objective and impersonal transfer of data and, as such, it is part of communication and represents a significant component of it, without, however, exhausting its scope. On the other hand, although conceptually distinguishable, communication and information are in a relationship of mutual implication, and not mutual exclusion. If, in fact, on the one hand, it is difficult, if not impossible, to achieve, through communication, reciprocity and the shared assumption of 'burden' without the 'exchange' of information, on the other hand, information, that does not take on the features of a gradual process within an articulated communicative relationship, appears inadequate and, in fact, impracticable, especially in the most problematic contexts and with the greatest impact on people's lives, and in the context of care relationships more than in any other.
Once one has, thanks to the clarification of the definition, a notion suitable for highlighting the value commitments and psycho-social dimensions inherent to communication, one can understand that its insertion in the care relationship is, first and foremost, functional to the valorisation of the different spheres of autonomy, the co-presence of which in the care relationship must be realised and guaranteed, without forgetting, however, that it is not a matter of autonomies that can be placed on the same level. And this, because of their different justificatory foundations, since the autonomy to be recognised and guaranteed to physicians is the autonomy justified by their professional competence, while the autonomy, in the strict 'decision-making' sense, to be recognised and guaranteed to patients, finds its justification in

the principle of the intangibility of the corporeal sphere and in the impact that interventions and treatments are destined to have on their health and on the quality, as well as the quantity, of their lives.
But communication, if is 'not misunderstood and taken seriously', can also play a key role in addressing and resolving the wide range of critical issues that the care relationship has to deal with.
The criticalities in question affect the care relationship as a whole, and concern both phisycians and care recipients. However, having regard to the aforementioned prerogatives and roles of both, a characterisation of them as 'criticalities on the side of patients' and 'criticalities on the side of physicians' can be proposed.
Among the former, mention must be made of the still too frequent unavailability of that information appropriate to the patient's condition, and consequently the absence, in the patient, of that adequate level of awareness about his condition and the impact of the activation or non-activation of treatments, in which the exercise of 'decision-making autonomy', in all contexts of care and, above all, in those of the highest criticality, has the necessary condition. A criticality, this, with which goes hand in hand that represented by the persistent reduction of informed consent to the bureaucratic fulfilment of the submission of a form, which the patient is required to sign. But further critical points, 'on the patient's side', both of which serve to fuel conflicting attitudes, are also to be found in the improper expectation of a therapeutic response and of a positive result that is always achievable (overestimation of the 'omnipotence' of medicine) and, on the other hand, in the conviction of being able to ask, with the legitimate claim of obtaining from doctors, any treatment, including those without evidence proving their efficacy A conviction, this, at the basis of which there is a dangerous misunderstanding of the notion of 'freedom of treatment'.
Among the 'critical issues on the side of physicians', there is, firstly, 'the knot of appropriateness', i.e. the difficult identification of appropriate and effective intervention strategies, thanks to which it is possible to achieve 'the best course of therapeutic action' in each specific case, in the face of the increased availability of alternative paths, as well as in the face of the diversification and expansion of the areas and purposes of medical interventions. Secondly, 'the knot of the right measure of treatment', i.e. the persisting resistance to therapeutic desistance and to the therapeutic switch to treatments aimed at controlling suffering, in the case of patients with a poor prognosis and/or in seriously disabling conditions, whether competent or no longer able to express their will.
With respect to this wide range of critical issues, communication, in the sense of interpersonal dialogue process between doctors and patients, if placed at the centre of the care relationship, and enhanced as its constituent element, can be the key to 'unbureaucratisation' of informed consent and the overcoming of obstacles to the patient's informed will, as well as to the remodulation of improper expectations with a predictable, as well as desirable, positive impact on the reduction of litigation. The establishment of an adequate communication process between doctor and patient can, on the other hand, facilitate the doctor in identifying the most appropriate therapeutic strategy, allowing him to refer to the patient's expected utility as a convenient, or even decisive, criterion to be adopted to select and propose treatments that maximise the probability of a favourable outcome. Moreover, in the case of patients suffering from pathologies with a poor prognosis, it is precisely the adoption of an articulated communication pathway that forms the basis of the care approach based on care planning, shared by the doctor with the patient, in which there are the prerequisites for end-of-life decisions that are capable of relieving the patients of suffering, respecting their wishes.

4. Medicine enriched by algorithmic tools. Communication as a route to risk mitigation

If, in the light of the foregoing considerations, we shift our attention at this point to the new scenarios of digitalised medicine supported by artificial intelligence applications, we must, first of all, emphasise how what characterises it and represents its distinctive element, and also its strong point, is, on the one hand, the extent of information relevant to addressing and solving health problems on which it can rely,

and, on the other hand, the availability of software capable of producing algorithms that provide answers or decision-making solutions, establishing correlations, extremely rapidly, between quantitatively extensive data, access to which would have been difficult, or rather impossible, for any doctor before the advent of the digital era.

Now, there is no doubt that information, amplified and accelerated by new technological resources, can play a key role in transmitting knowledge on which to base a good clinical decision. It would be wrong, however, to underestimate the risk, which has already been mentioned in the literature on the subject[8] and is also being considered in the institutional sphere[9], of the inadequacy (reliability) of the data at the basis of the algorithms destined to find application in the health sphere, due, for example, to the insufficient presence in the datasets of data relating to certain populations or to the failure to take into account individual differences existing within a population, but also to errors of measurement and classification.

Even if one were to intervene on such biases at the design stage, so as to have fairer and more reliable algorithms[10], one will not, however, have tools that can replace the doctor in the task that is proper to him, that, as noted above, of identifying and proposing treatment paths appropriate to the patients' conditions. In order to fulfil this task, without indulging in conformist solutions, adopted by flattening on acquired data, the integration of algorithmic indications with the anamnesis and the objective examination will remain the way to go, if one does not want to overlook individual peculiarities, which escape categorisation, and on which a good diagnostic and therapeutic framing often depends.

Thus, the information processed by artificial intelligence is being brought back into the care relationship as a two-way communicative context, dialogic and not monologic, as well as marked by empathy and trust, in which there is still room for some additional anamnestic information, relevant for diagnosis and therapeutic choice, but previously unavailable or not considered. It must, on the other hand, be emphasised that, in the pathway aimed at identifying the best treatment strategy, of what is to be done in the individual case, it is not enough to have knowledge based on empirical data, not even if the knowledge in question is that derived on a large amount of data, and we might add, of quality data, the correlations of which artificial intelligence software reliably highlights.

If one reasons in terms of the relationship between means and results (or, if one prefers, the relationship between means and ends), one must, in fact, recognise that science, even algorithmic science, can certainly already now, and even more so in the future, contribute to defining possible intervention scenarios and to envisaging achievable results. Neither now, nor in the future, can it, however, say anything about the merits of the results[11]. This is where evaluations come into play, the shared principles of which at the bioethical level, and also at the legal level, have recognised that it is up to the person, whose health and life are at stake, i.e. the patient, whose understanding of expectations, needs and values can play a leading role in defining therapeutic appropriateness itself.

Once we have focused our attention on the patient, whose involvement in the clinical decision-making process and whose decisive role in treatment decisions is no longer in question in terms of ethical, deontological and legal principles, it remains, at this point, to emphasise that the possibility of autonomously accessing health contents processed by artificial intelligence software, and present in digital platforms, is not facilitating the patient's acquisition of information functional to the adoption of informed decisions. On the contrary, it sometimes generates the unfounded conviction of having found, already defined in its contours, the therapeutic response that is well suited to his specific case, and whose implementation can be demanded by the physician. Or, on the contrary, it arouses equally prejudicial opposition, especially when the patient encounters the prospect of therapeutic strategies that are highly innovative with respect to the traditional ones that have long been known and practised. Faced with the risk, on the one hand, of uncritical reliance, as a result of an attitude of overreliance on the 'machinery', in the diagnostic-therapeutic field as well as in the case of robotic instruments in the surgical field, and on the other hand, of a no less uncritical rejection, due to distrust of what one does not know how it

works, there does not seem, at present, but also in the future, to be any alternative to that of establishing an in-depth dialogue with the patient. This is the context in which the physician can re-dimension the idea of self-sufficiency, or, at least, of the necessary prevalence of the algorithm over further informative elements that have emerged in the confrontation, which are important in order to clearly envisage the therapeutic alternatives that can be proposed and can to support the patient in an informed choice. And it is, more generally, the context in which one can help to reduce the understandable disorientation aroused by the so-called 'opacity' of the algorithms, that is, the difficulty, or impossibility, of understanding the computational processes that lead to certain results through data processing.
The idea that applications of artificial intelligence, in the field of health, even more than in other fields, should not be separated from the recognition of the 'right to an explanation', so as to understand 'the reasons and circumstances that led to a specific decision taken by the algorithm'[12], is becoming increasingly rooted on an ethical level. But for this right to be guaranteed, it is necessary that the doctor first be put in a position to understand artificial intelligence systems, being involved in the process of designing and understanding the datasets that can influence the decision on treatment. This is a journey that is only at the beginning and that will present the physician with demanding training challenges. Achieving it will be important not to rethink the care relationship that has its constituent element in communication with the patient, but to enable its fullest implementation in the face of new scientific and technological frontiers, which it is as necessary to govern as it would be absurd to try to stop.

Bibliographical references

1. Borsellino P., Bioetica tra “morali” e diritto. Nuova edizione aggiornata. Milano: Cortina editore; 2018, in particular chapters three and four.
2. Council of Europe, Convention on Human Rights and Biomedicine, Oviedo 1997; UNESCO, Universal Declaration on Bioethics and Human Rights, 2005; World Medical Association International, Code of Medical Ethics. 2022; FNOMCEO, Italian Code of Medical ethics, 2014.
3. Italian Parliament, Act n. 219/2017 “Rules on informed consent and advance treatment directives”, art. 1.1.
4. Berger R., Bulmash B., Drori N., Ben-Assuli O. et al. , The patient–physician relationship: an account of the physician’s perspective, Israel Journal of Health Policy 2020; 33: 1-1.
5. Italian Parliament, Act n. 219/2017 “Rules on informed consent and advance treatment directives”, art. 1.8. There, communication time is qualified as care time.
6. Collecchia G., De Gobbi R., Intelligenza artificiale e medicina digitale: Una guida critica, Roma: Il pensiero scientifico italiano; 2020.
7. Salardi S., Intelligenza artificiale e semantica del cambiamento: una lettura critica. Torino: Giappichelli editore; 2023, p. 118.
8. Cabitza F., Rasoini R., Gensini G.F., Unintended consequences of machine learning in medicine, Jama 2017; 318 (6): 517-518
9. Council of Europe, Steering Committee for human rights in the field of biomedicine and health (CDBIO), The impact of artificial intelligence on the Doctor-patient relationship, 2021.
10. As called for by UNESCO, Recommendation on the ethics of artificial intelligence, 2021.
11 Borsellino P., Conoscenza e diritto La prospettiva della riflessione filosofico-giuridica di orientamento analitico, Rivista internazionale di Filosofia del diritto 2022; 1-2: 19-38.
12. Luverà C., Il machine learning come strumento di supporto nelle decisioni mediche: questioni etiche e prospettive, Bioetica. Rivista interdisciplinare 2021; 3: 422.

Ethics and Artificial Intelligence in the Doctor-Patient Relationship"

Rosagemma Ciliberti; Linda Alfano
Department of Health Sciences, University of Genoa, Genoa, Italy

The use of Artificial Intelligence (AI) in the healthcare field can offer multiple and conflicting contributions. These technologies can enable healthcare professionals to reduce the time required for routine bureaucratic tasks, which can sometimes be sterile and divergent from the interests of the patients, and allow for increased opportunities to listen to the patient and enhance the quality of the caregiving relationship. On the other hand, this automated cognitive assistance can also diminish or undermine the relational skills and abilities of healthcare personnel.
For these reasons, the impact of AI on clinical care and the doctor-patient relationship calls for the need to develop shared ethical criteria to protect patient autonomy, ensuring transparency, equal opportunities, privacy, and security, all while promoting training programs for healthcare professionals not only in the technological domain but also in the ethical and social aspects. This should also include the incorporation of ethical discussions in the education of engineers, computer scientists, and developers, with specific emphasis on the ethical implications of technology design and its application to human beings. Additionally, it is important to foster an increasing awareness among the general population regarding the opportunities and risks associated with new technologies.

Introduction

In the past, the work of a doctor was primarily centered around the patient's bedside and involved direct manipulation of the patient's body, as well as the assumption of all responsibilities and decisions regarding their health by the healthcare professional. It was almost always the trusted physician who occupied a central and absolute reference point for the patient and their entire family.
The essence of this relationship lay in the interaction and the doctor's ability to positively engage with the patient, to intuit the nature of the illness and its treatment through careful observations of the individual in their various facets: posture, skin color, odors, non-verbal gestures, mood, and the management of daily life. Each encounter between the doctor and their patient constituted a kind of ritual that solidified the patient's trust in the doctor, owing to the physician's greater mastery of knowledge regarding the structure and functioning of the body.
In more recent times, with the evolution of medicine, the conventional doctor-patient relationship has gradually evolved into a composite and multidisciplinary relationship between a team of professionals and an individual in need of care, who also holds rights and the ability to exercise them. The traditional practice of a medical visit, characterized by empathy, human understanding, direct visual and physical contact, has increasingly been replaced by the examination of results from mechanized and computerized diagnostic techniques, such as images displayed on screens, X-rays, medical reports, numbers, and statistical information. These are high-quality tools but not always capable of capturing either the holistic reality of the individual or their most authentic needs.
Profound social, cultural, and ethical transformations have also long diminished the centrality that the traditional doctor traditionally held in their role as the exclusive custodial carer.
The emergence of the phenomenon known as 'web 2.0' has introduced further innovative elements into the relationship between doctors and patients. Today, medical authority no longer represents an unquestionable certainty. The citizen, increasingly impatient, has access to a wide range of online resources, more or less reliable, that influence their decisions, sometimes creating distortions, unwarranted alarm, and illusory hopes. This change in the doctor-patient relationship reflects a broad

cultural and political transformation that has redefined, expanding the informational and decision-making power of every individual. Healthcare professionals must face the challenge of striking a balance between opening up to the active participation of patients in managing their own health, with the potential for challenges or outright rejection, and the role of guidance and advice they claim based on their training and expertise.
The proliferation of AI introduces further new actors capable of improving the efficiency and accuracy of medical practices and, at the same time, significantly impacting not only the ways in which medicine is practiced but also the relationships between the various parties involved in an increasingly complex scenario. Indeed, there are multiple applications of AI that encompass diagnosis, treatment, prevention, patient monitoring, and rehabilitation pathway control. These applications also extend to the clinical trial phase and involve management and supervision operations within the complex healthcare system, improving its functionality, accessibility, and reducing costs related to health data management and medical records.

In the face of what are now commonly considered irreversible and unstoppable developments, which have led to significant successes and new opportunities, concerns and risks have also emerged. These necessitate a thorough examination and careful reflection on the ethical implications involved, encompassing both the new roles of doctors and various healthcare professionals in an increasingly technologically advanced environment and the possibility of maintaining medicine in a human dimension that continues to preserve its foundational elements of empathy, listening, and dialogue.

The Challenge of AI
In order to further delineate the risks and challenges associated with the use of AI, in addition to the previously mentioned issues, it is important to focus on a key concept that permeates and characterizes the medical profession: trust.
Trust plays a fundamental role both within the relationship with the patient and in the broader context of AI implementation. It involves the trust that doctors place in AI-based tools and also the trust that patients place in the decisions and recommendations coming from these tools.
This trust, an increasingly critical dimension in the current healthcare landscape, which is becoming more efficient yet also increasingly distant from the needs of the individual, encounters further challenges in the realm of AI due to the intrinsic nature of the system itself.
These devices, capable of continuous adaptation and learning, are built using internal mechanisms that are not easily interpretable and explainable. AI does not follow a linear, predefined, or predictable path through a software of algorithms; rather, it adopts a self-learning approach based on the machine's own past experiences or those acquired from the surrounding environment, as seen in the case of 'Deep learning.'
This opacity phenomenon, known as the 'black box,' poses ethical issues, especially when it comes to critical applications like healthcare. Algorithms can produce accurate results, but often there is a lack of clarity on how they arrived at these conclusions, making it impossible to understand the logical steps and considerations that led to these decisions, even by those who designed them. Additionally, algorithm biases that arise during the machine's training phase in relation to data selection and methods can further exacerbate these interpretative deviations and errors. These issues are particularly relevant in the medical context because the use of intelligent systems can pose risks to the health and lives of patients.
This characteristic of inaccessibility to the 'internal reasoning' of the device implies a limitation in understanding and verifying the entire process, which can undermine the overall reliability of the system. Given the inability to fully comprehend the decisions generated by AI, the doctor may be compelled to relinquish their role as the primary driver of care dynamics, both in terms of determining the treatment pathway and in guiding the patient towards informed and responsible healthcare choices.

Furthermore, it should not be underestimated that the doctor and AI may arrive at divergent diagnostic conclusions, creating a dilemma regarding the choice to be made. For instance, let's assume that AI, based on data, suggests that a medical condition requires immediate surgical treatment, while the doctor may prefer a more conservative approach. In this situation, the opacity of the reasons guiding AI could raise ethically significant questions concerning decision-making responsibility and patient safety, as well as the proper application of ethical principles of beneficence and non-maleficence.

As emphasized by the Italian National Bioethics Committee, the unintelligibility of the process by which an AI system arrives at a specific diagnostic or therapeutic option could inhibit the doctor from making autonomous assessments different from those suggested by the machine, in the name of a presumed 'technological superiority.' Moreover, an unquestioning attribution of decision-making priority to the machine would not only lead to dangerous dogmatic tendencies but also a resurgence of the paternalism paradigm, albeit shifted from the figure of the doctor to the symbolic one of AI. As highlighted, opposing a scientifically authoritative decision could prove complex and require challenging emotional and cognitive autonomies that are not always readily available.

Regarding this, one should remember the social psychology experiments conducted in 1961 by the American psychologist Milgram (1) on obedience to authority, specifically to a scientist who ordered participants to carry out actions (administering electric shocks to hypothetical students) conflicting with their ethical and moral values. Contrary to expectations, a significant percentage of people displayed a surprising degree of obedience, even in violation of their moral principles. Milgram believed that obedience induced by a legitimate authority figure could be attributed to a kind of heteronomous state characterized by the fact that the subject no longer considered themselves free to undertake autonomous actions but merely instruments to carry out orders. The experiment's subjects did not feel morally responsible for their actions but rather as executors of the will of an external power.

The creation of this 'heteronomous state' is influenced by three factors: the perception of authority's legitimacy (in this case, the experimenter embodied the authority of science, as could potentially happen with AI); obedience education as part of socialization processes; and social pressures (for our discussion, the myth of the omnipotence of science and technology combined with the taboo of death and its denaturalization).

Subsequent research that utilized Milgram's paradigm, such as that conducted by David Rosenhan, confirmed the results obtained by Milgram but also highlighted that the degree of obedience to authority varied significantly based on the physical and emotional distance between the experimental subject (in the role of the teacher) and the experimenter (in the role of the scientist). Four levels of distance between the teacher and the learner (an actor specially trained) were tested: in the first, the teacher could neither see nor hear the victim's complaints; in the second, the teacher could hear but not see the victim; in the third, the teacher could hear and see the victim; in the fourth, to administer punishment, the teacher had to physically hold the victim's arm and push it onto a plate. In the first level of distance, 65% of subjects went on to deliver the strongest shock; in the second level, 62.5%; in the third level, 40%; in the fourth level, 30%.

These results, if applied in the healthcare context, once again underscore the importance of the doctor-patient relationship to protect all participants in the care pathway.

Milgram also demonstrated that obedience depended on the underlying ideology that defines and explains the meaning of events and provides the perspective through which individual elements gain coherence. The likelihood of engaging in certain behaviors over others is influenced by an individual's perception of the situation, which determines which norms are relevant to the context and therefore should be followed. In other words, if the subject (doctor) accepts AI's definition of the situation, they may end up accepting the proposed actions, even if questionable, not only as reasonable but also as objectively necessary.
These considerations should, in healthcare contexts, suggest special attention to the respect, dissemination, and sharing of fundamental ethical principles such as respect for the individual, vulnerability, and responsibility.

Delegating decision-making to the machine does not only result in a reduction in human attention but also a diminishing role of the doctor in the decision-making process in favor of AI. This choice could lead to a dangerous phenomenon known as 'deskilling,' which involves a loss of capabilities and dequalification linked to an over-reliance on mechanical output, capable of generating the described psychological mechanism of deresponsibilization regarding the recommendations made by AI.
Delegating medical decisions to automated systems could induce a sort of professional passivity on the part of the doctor, with a potential reduction in their ability for clinical discernment and evaluation of each patient's unique situation. This could also create a disconnect between the healthcare professional and the patient, with legal consequences that need to be carefully considered.
Clearly, the issue of 'to whom' to entrust oneself doesn't only concern the healthcare provider but also involves the patient. In the absence of a clear perspective on a reliable diagnostic and therapeutic scenario, patients may have difficulty understanding and evaluating the motivations that led to a specific proposal.
Informed, free, and responsible consent forms the ethical foundation, even before it becomes a matter of professional ethics and legality, for the therapeutic relationship. This is essential to counteract the potential 'dehumanization' of the therapeutic relationship, which could lead to a loss of the fundamental elements of dedication to the uniqueness and individuality of the patient, even before considering their illness.
Another equally important aspect to consider is the risk that the algorithm may prioritize available options based on a hierarchy of values that do not align with the patient's cultural, anthropological, and existential principles. A concrete example could be the algorithm's orientation towards a treatment path that prioritizes a longer life expectancy, favoring quantity of time over quality, while the patient may have opposite preferences. Similarly, algorithms could steer towards medical practices that satisfy administrative or economic objectives rather than the patient's actual care needs.
Furthermore, we cannot overlook the possible paradoxical effects that might lead some people to reject new treatment methods, while others may give consent that is not sufficiently considered or responsible. These issues related to the explicability of the decision-making process have also raised the hypothesis of a denialist attitude, which could exempt the healthcare professional from the obligation to inform the patient about the use and mode of implementation of such technology in the case of AI utilization.
The ability to provide genuinely informed and comprehensive information in a 'black-box' context is undoubtedly a complex challenge that requires the implementation of specific and appropriate regulatory tools. It also demands significant effort in terms of dialogue and collaboration among various professions and expertise, including bioethicists, legal experts, programmers, computer scientists, developers of new technologies, and risk assessment specialists, in addition to healthcare professionals.

This commitment is essential to avoid the risk of 'dehumanization' of the care relationship, meaning the danger of losing the elements of attention directed towards the uniqueness of the assisted person.

In this context, it is appropriate to refer to the document 'Artificial Intelligence and Medicine: Ethical Aspects' issued by Italian National Bioetichs Committee and Italian Committee for Biosafety, Biotechnology and Sciences of Life (CNB, CNBBSV) on May 29, 2020 (2). This document, recognizing the challenges associated with obtaining informed consent in the use of AI, explicitly states that 'it is an ethical and legal obligation that those undergoing such innovative healthcare treatments through AI are informed in the most suitable and understandable manner about what is happening, whether they are (if applicable) subject to experimentation and validation, and that they are aware that what is being applied to them (diagnostically or therapeutically) entails advantages but also risks. It should be explicitly specified in the informed consent whether the treatments applied (diagnostic or therapeutic) come solely from a machine (AI, Robot), and whether and what the scope and limits of human control or machine supervision are.

In line with this approach, the report 'Artificial Intelligence Systems as Diagnostic Support' by the Ministry of Health (2021) emphasizes the indispensable need to extend the scope of information regarding the use of new tools, while also providing specialized training for both doctors and patients (3).

Undoubtedly, the inherent challenges in the AI system require great attention in defining the boundaries of information, which should be flexible, calibrated, and modulated, but should not be completely excluded (4).
A clear indication in this regard comes from Italian Law 219/2017 'Provisions on Informed Consent and Advance Treatment Directives,' which establishes that this information must be part of a relationship of 'care' and 'trust (5), and can even become a tool of care itself, regardless of whether the healthcare treatment is 'intelligent' or not.
This emphasis on the relationship of care and trust cannot be entirely delegated to the intelligent system but must remain under the direct control of the healthcare professional to preserve the crucial role of the specificity of this relationship.
The statement contained in Law 219/2017 that 'The time for communication is a time for care' represents a clear guideline in this regard. This principle underscores that the moment of communication between the doctor and the patient is not just a formal step but a crucial moment in which trust is built and nurtured, information is shared, and a shared and responsible decision-making process is ensured.

Conclusion
Effectively addressing the diverse ethical issues arising from the development of AI in medicine is a challenging task that requires abandoning "individual" and "fragmented" approaches in favor of comprehensive and articulated strategies based on constant dialogue among diverse knowledge and expertise.

In an editorial published in the New England Journal of Medicine, Ziad Obermeyer and Thomas Lee emphasize the limits of the human mind compared to the current complexity of medicine, the vast amount of existing data, but also the need for analysis, contextualization, and interpretation. This is a daunting task if we continue to rely on past methodologies, such as simple discussions among multiple doctors about the clinical situation and the subsequent solution to adopt (6).
The first step to take is to recognize the disparity between the abilities of the human mind and the complexity of modern medicine.

In this regard, a thorough reconsideration of the training of healthcare professionals is important, aimed not so much at adding missing knowledge, but rather at epistemologically reforming the available knowledge through an approach capable of weaving relationships between various disciplines. In this context, the reflection of Alfred North Whitehead in his work "Science and the Modern World" from the now-distant 1926 seems particularly appropriate (7). The great mathematician and philosopher warned of the pitfalls arising from hyper-specialization, arguing that while it increased the sum of knowledge in specific fields, it negatively affected the realm of knowledge, producing one-dimensional minds incapable of comprehending the complexity of circumstances. This complexity expresses the interconnection of systems and allows properties not possessed by individuals but only by their interaction to emerge.

This rethinking in the training of medical professionals, as previously mentioned, should also extend to other related technical professions such as engineers, computer scientists, and developers.
Strategies and policies that break down the barriers between knowledge and expertise are desirable, especially in the management of a technology that, currently used in a limited way compared to its potential, is destined to increasingly change the essence of society and medicine in the near future.

What is needed is the development of a different cultural approach that can reconcile the relationship between technology and healthcare needs, in order to make use of high-value tools that are oriented towards the real needs of individuals and the respect for dignity and freedom of choice.

References

1. Milgram, Stanley. (1974), Obedience to Authority; An Experimental View. Harpercollins (ISBN 0-06-131983-X).
2. Presidenza del Consiglio dei Ministri, CNB, CNBBSV “Intelligenza artificiale e medicina: aspetti etici” del 29 maggio 2020. https://bioetica.governo.it/media/4260/p6_r_2020_gm_intelligenza-artificialeit.pdf
3. Ministero della Salute. I sistemi di intelligenza artificiale come supporto alla diagnostica", 9 novembre 2021. https://www.salute.gov.it/imgs/C_17_pubblicazioni_3218_allegato.pdf
4. Astromske K. Peičius E, Astromskis P. Ethical and legal challenges of informed consent applying to AI in medical diagnostic consultations, in AI&Society, 36, 2021, 509-520.
5. Ciliberti R, Gorini I, Gazzaniga V, De Stefano F, Gulino M. The Italian law on informed consent and advance directives: New rules of conduct for the autonomy of doctors and patients in end-of-life care. Journal of Critical Care 2018; 48: 178-182.6. Obermeyer Z, Lee TH. Lost in Thought — The Limits of the Human Mind and the Future of Medicine. N Engl J Med 2017; 377:1209-1211.
7. Whitehead AN. (1997) La scienza e il mondo moderno. Free Press

AI and Mental Health: new challenges from the Ecotechnobioethics

Prof. Dr. Moty Benyakar M.D.; Ph.D.; Prof. Lic. Nicolas Obiglio

Department of Bioethics and Artificial Intelligence, International Chair in Bioethics.

Introduction

In the face of technological advances, particularly in artificial intelligence (AI), humans confront with new challenges. Technology leads us to new models of human interaction, forms of expression of will, and a change in the way we relate to each other, as well as in the modalities of seeking our well-being. During the COVID-19 pandemic, humanity was forced to socialize through virtual spaces. This type of interaction has led to different modes of human relationships; for example, being in contact with loved ones without physical contact already poses a new modality, just as the development of telemedicine raised crucial questions about mental health.

The proliferation of tools like chatbots based on deep learning, such as ChatGPT, has led AI to make an impact, modifying fundamental areas of human life internationally, from the economy to education, and especially mental health. However, we could assume that these technological advances have impacted human subjectivity, raising questions about its meaning of its development.

In response to these changes, Ecobioethics must take a proactive role to anticipate and safeguard the bioethical and fundamental principles of human beings. In the context of the advances of emerging technologies, challenging our understanding of their impacts on human subjectivity, it is crucial to redefine the role of Ecobioethics. We propose the use of the term "Ecotechnobioethics," as a broad conception of the impacts of technology, that includes the Ecobioethics; that addresses the impact of AI development on human psyche, with the aim of sustaining the development of human subjectivity while respecting principles such as beneficence, autonomy, prevention of harm, among others.

1. The Development of Ecobioethics: Towards Ecotechnobioethics

Twenty-five years ago, Moty Benyakar initiated the exploration of Bioethics, revealing that this concept was limited to the perspective proposed by Van Rensseler Potter, a biochemist and oncologist who, in 1970, redefined the term Bioethics, focusing primarily on patient rights and the principles of the doctor-patient relationship.

Subsequently, the true creator of the Bioethics concept was the German pastor and theologian Fritz Jahr (1895-1953). Jahr conceptualized Bioethics as a Global Ethics, proposing the idea of a "Universal Bioethical Imperative" to replace Kant's formal imperative.

Faced with this conceptual restriction, Benyakar proposed the concept of Ecobioethics to UNESCO, where doctor-patient relationship is seen only as one dimension of human relationships. The Eco addresses to Bioethics goes beyond the ecological perspective, considering not only the interaction between humans and the physical environment but also extending to what lies beyond human behaviour (1).

These developments led to in-depth investigations, subsequently presented to UNESCO. Currently, within the framework of the International Bioethics Chair (ICB), the Ibero-American Network of Ecobioethics has been established, presided over by Moty Benyakar and Rui Nunes (2).

This concept is fundamental because it opens the door to the study and research of Bioethical factors in different areas of interaction between humans and technology. Drawing from the Complexity paradigm proposed by Edgar Morin, and in the face of cybernetic developments, we can investigate aspects such as coexistence, well-being, and health, as well as psychological, social, educational, and cultural factors, among others. Thus, we can delve into the human interaction with artificial intelligence and mental health.

In the context of our research on the impact and interaction between humans and artificial intelligence, based on Nicolas Obiglio's proposal, we introduce the neologism "Ecotechnobioethics." This new branch of Ecobioethics specializes in addressing advances in technology, cybernetics, and especially AI's impact on human subjectivity. The purpose of Ecotechnobioethics is to focus on the interactive developments of AI and human subjectivity, both in its beneficial use and its potential harmful impacts. The goal is to safeguard the mental health of individuals and protect fundamental rights in the face of advancing AI technologies.

2. Justification: Impact on the Psyche

What is the role of artificial intelligence (AI) in our existence? It is crucial to understand that we are not merely in relation to technology; rather, we consider it inherent to our own subjectivity. We do not conceive technology as internalized, but rather as actively participating in our processes of internalizing the factual. (3)(4)(5)

Mind and Psyche

To address this theme, we propose a conceptual distinction between the mind and the psyche. While we understand that both are in constant interaction, and their description is purely functional, the mind or the mental is a product of our brain activity, whereas the psyche is a product driven by the drive that is under the aegis of the unconscious.

In the mind is where human intelligence develops; it’s a phylogenetic capacity, just as animals posses’ intelligence. Both us and them, we develop our intelligence through our brain, meaning it’s one of the mind’s characteristics. On the other hand, the Psyque is constituted by our human capacity to represent, that is, the subjective transformation of what is presented to humans. That is the idiosyncrasy of humans. Therefore, we understand that AI replaces or aims to replace human intelligence but not the subjective transformation produced by the Psyche. (8)

On the other hand, the difference lies in that the mental confronts what is, and the psyche deals with what is not or what can be for each individual. In other words, the mind operates within the emergence of concrete phenomena, while the Psyche operates with symbolic dimensions. (3) (4) (5)

However, human essence does not reside solely in the ability to think but in the ability to represent, which constitutes us as idiosyncratic subjects. This capacity involves

transforming the factual[1] into the psychic, allowing the psyche to translate reality and represent it, thus creating a unique dimension for each subject in their way of representing the factual. (3)(4)(5)

AI emulating intelligence, not the psyche

Artificial Intelligence emulates mental function, not the psychic and symbolic dimension. We use the term "emulation," which comes from the Latin "aemulare," meaning to strive to equal or surpass, suggesting competition and effort to achieve a level of excellence. Although AI far surpasses human mental capabilities, it cannot replicate the psychic dimension, namely the representative function that is an exclusive quality of the human being. (6)

On the other hand, AI performs transformations of what is presented through algorithms or codes, similar to our body cells, but never psychic transformations. The crucial distinction is that AI can artificially develop thought but not affection and representation. Although AI emulates words and functional thinking, as seen in the case of ChatGpt or some cognitivist techniques using AI functions to elaborate cognitive thoughts, it lacks a real understanding of the content it expresses.

In the linguistic realm, there is a significant difference. The meaning of words is the mental correlate of the signifier. For the letters not to be mere linguistic signs, there is a psychic activity that has a symbolic component, i.e., the dimension of the signifier. Artificial intelligence constructs sentences and texts based on algorithms, predicting the next word based on its training and learning, but it does not understand its subjective meaning.

The symbolic dimension of the psyche and factual dimension of the mind

In the symbolic dimension, a distinction is made between the signifier and meaning to emphasize that the meaning of words is not fixed but subject to changes and transformations. The meaning of a word can varies depending on the context in which it is used or the speaker's intention. (7)

The fundamental difference lies in the absence of the deployment of the symbolic dimension in AI, which is inherent to human subjectivity. AI achieves a mentalization or mathematization of the meaning of words. The use of AI words refers to a meaning but not to the symbolic dimension; it learns the meaning of words through trial and error, through a binary structure between what is and what is not, as a problem-solving approach to what is presented.

Humans transform the factual into representations through psychic activity, a transformative process that each psyche undergoes. We assert that humans possess an exclusive capacity for idiosyncratic representation, rooted in unconscious processes that establish our differences and subjectivity. (3)(4)(5)

---

[1] The factual is what is, in disciplines separate from psychoanalysis, we could say that it is reality. Therefore, the essential human aspect is the capacity to psychize reality, that which is presented to us, to be represented by our psyche.

3. Ecotechnobioethics: In the Human Bio-Psycho-Social Dimension

After clarifying the primary difference between the mind and the psyche, the key challenge of technological development is that it brings about a change in our way of internalizing the factual. In other words, humans, when interacting with the environment, transform both their surroundings and themselves. In the ethereal era, technology created by humans leads us to new models of human transformation, new ways of relating and experiencing the world, as well as seeking well-being.

Ethereal means that is intangible or loosely defined, guide us towards an understanding of our contemporary era, which Moty Benyakar has label as Ethereal Era; due to the rapid, amorphous, intangible transformation, not always perceptible, where everything is everywhere, and simultaneously nowhere. (8)

The development of AI impacts all spheres of the human being, understood as a bio-psycho-social being. One of the functions of Ecotechnobioethics is to observe and anticipate the impact on the psyche, safeguarding human subjectivity in social[2] dimensions, preserving the humanization of human connections, and biological[3] dimensions.

4. Analysis of a Technological Phenomenon: "Digital Necromancy"

This is an example of the application of Ecotechnobioethics and its inquiry in regarding human subjectivity. In this analysis, we will delve into the study of an emerging technology with the potential to influence our psyche: Digital Necromancy. In order to preserve the mental health of individuals, we focus on understanding the impacts that this technology could have. (9)

Digital Necromancy involves the use of artificial intelligence to create digital simulations of deceased individuals. These simulations can be used to interact with the deceased through conversations, games, and even more complex tasks. Although still in development, there are already companies, such as the American Eternime, offering services in this area, allowing people to create digital avatars that are updated with their online content after their death.

Digital Necromancy could generate a series of impacts on the human psyche that become ambiguous:

- *Consolation for mourners*: Digital emulations of deceased individuals could help subjects process the loss and maintain a connection with the deceased. However, it could also have the opposite effect, posing a danger to the grieving process, resulting in various pathologies that directly affect the mental health of individuals, leading to a deterioration in their quality of life. (10)

- *Education*: Digital simulations of historical figures could be used to educate future generations about their lives and contributions. However, they could also be used to manipulate information and respond to political interests.

---

[2] As it has been doing, in the use of new virtual communication spaces, for example.
[3] The use of AI in the field of health. (detection of diseases, new treatment modalities, etc.)

- *Creativity:* Digital simulations of famous individuals could be used to create new forms of art and entertainment. For instance, emulations of the voice through AI, featuring current songs sung by deceased artists. However, this could lead to numerous legal conflicts regarding intellectual property, such as the voice, or even compel deceased singers to create music that they would never have done in life. Hence, governments must begin regulating such creations.

- *Deception:* Digital simulations of deceased individuals could be used to deceive the living into believing that they are interacting with a real person. Additionally, there is the possibility that the algorithm makes the deceased say things they never would have. This could have a negative impact on the mental health of people who have lost a loved one.

- *Addiction:* Digital simulations of deceased individuals could be addictive, leading people to lose touch with reality and develop pathologies such as social phobias, causing a deterioration of mental health.

These impacts pose a series of problems that Ecotechnobioethics must address, including issues of privacy, intellectual property, and consent. Digital Necromancy challenges the sense of finitude present in existentialist currents by preserving the voices of the deceased and raising questions about legal responsibility and intellectual property. So far, the use of the voice[4] is not regulated and should be preserved. The debate arises to introduce a new ethical dimension regarding the intellectual property of the deceased and their environment, the family. There may be family members who do not wish the voice of the deceased to continue to be used, and if they do, who would be responsible?

These complexities demand the intervention of Ecotechnobioethics to collaborate with governments in establishing rules that regulate the use of this technology for the benefit of our own subjectivities.

5. Ecotechnobioethics

Therefore, we will propose some initial principles belonging to what we understand as Ecotechnobioethics, as the basis for future developments: (11) (12)

1. *Human Rights:*

- *Right to Life, Liberty, and Personal Security*: AI must respect and protect the right to life, liberty, and personal security. AI applications, especially those related to critical decision-making, must be designed not to compromise the safety and well-being of individuals. Safeguards must be established to prevent situations that endanger the lives or freedom of people.

- *Equality and Non-Discrimination*: AI algorithms can inherit and perpetuate social biases, leading to discrimination. It is imperative to ensure that the implementation of artificial intelligence does not exacerbate existing inequalities or discriminate against

[4] In contrast, the use of images is mostly regulated; the use of body images or photos of individuals without authorization or of children, even with technologies like Deep Fake, poses a real danger to privacy and subjectivity.

individuals or groups. Principles of fairness and non-discrimination must be incorporated into the design and evaluation of algorithms to avoid harmful outcomes.

- *Freedom of Thought, Conscience, and Religion:* Technologies should refrain from coercing individual beliefs. The implementation of AI must respect the diversity of thought and ensure that algorithmic decisions do not limit freedom of expression of opinions or religious beliefs. Additionally, special attention should be given to the new subjectivities that may arise from interaction with AI or technology. Moreover, the emergence of absolute answers from AI may lead to a detachment from the spiritual dimension of each human being.

- *Freedom of Expression:* The development of artificial intelligence must protect and promote freedom of expression. Algorithmic censorship and information manipulation must be avoided. Individuals should have the ability to express their opinions freely without fear of automatic reprisals. Diversity of voices and perspectives must be preserved and respected. Additionally, there should be a special emphasis on protecting ourselves from interpreting AI-generated responses as absolute judgments of truth or falsehood, as a conflict between absolute knowledge and more relative postulations.

- *Right to Work and Fair Working Conditions*: AI-driven automation can impact employment and working conditions. It is essential to safeguard the right to work and ensure that the implementation of AI contributes to the creation of meaningful employment and fair working conditions. Retraining and workforce adaptation should be priorities to mitigate negative impacts.

2. *Justice:* AI must avoid biases and inequalities. Moreover, it is important to promote the inclusion and education of all communities about the development and use of AI to prevent inequalities. (13)

3. *Autonomy:* Individual autonomy is a central principle that must be safeguarded in AI development. Algorithmic decision-making must be transparent and understandable to individuals, allowing them to maintain significant control over decisions affecting their lives. The application of AI should empower people, facilitating informed decision-making and respecting the diversity of perspectives and values. This aims to avoid using information to manipulate decisions that individuals may make in using AI. (14)

4. *Beneficence:* This principle emphasizes to ensure that AI benefits society as a whole. Developers and companies should direct their efforts toward creating technologies that improve the quality of life, promoting equity and inclusion. Considering the impact it could have on human relationships, extreme precautions should be taken with subjectivities that are at risk. Solutions should be sought to trying to find solutions to face social problems and contribute to collective well-being, minimizing disparities and improving accessibility. (15)

5. *Veracity:* Transparency and truthfulness are essential in the interaction between humans and artificial intelligence. Clear communication about how AI operates, how data is collected and used, and potential impacts on decision-making is essential. Additionally, it is imperative that humans be informed when interact with AI and not with a human. For this reason, one of the goals is to develop cybernetic tools that enable these differentiations. (16)

6. *Confidentiality*: Confidentiality in the context of AI relates to the protection of privacy and data security. Ecotechnobioethics principles require robust safeguards to ensure that sensitive information is handled securely. Users must be assured that their personal data will not be used inappropriately or shared without their consent. Transparency in the acquisition of data by AI and the organizations using it is also crucial. (17)

7. *Awareness of Misuse*: Non-maleficence requires minimizing the harms and risks associated with the implementation of AI. It is essential to identify and mitigate potential negative consequences, impacts on new forms of human subjectivity, algorithmic biases, discrimination, and job loss. Ecotechnobioethics should assess potential impacts and develop preventive measures to avoid unnecessary harm. (18)

8. *Ecotechnobioethics in Technological Development*: The pursuit of benefit must go hand in hand with an ethics of technological development, where innovation does not compromise the integrity of human subjectivity, and the risk of collateral damage is minimized.

9. *Ecotechnobioethics in Arts Development*: The use of AI for art creation, safeguarding art as a subjective manifestation and not merely mechanized; also, the use of technology to emulate deceased artists. (19)

10. *Prevention of Psychic Harm*: Given the uncertainty about the psychological impacts of AI, Ecotechnobioethics advocates for preventive measures, promoting research and regulation to protect mental health in the use of these technologies.

These principles of Ecotechnobioethics serve as a guide for the development of artificial intelligence, ensuring that technology enhances human life without compromising subjectivity, autonomy, and fundamental rights. Their diligent implementation is essential to build a future in which AI coexists harmoniously with the human essence and the spontaneous continuity of each individual.

Furthermore, Ecotechnobioethics places special emphasis on the research and development of AI use that is beneficial to humans. The unquestionable positive impact it has generated and can generate in many areas of human life, such as science (accelerating research processes, improving data accuracy), healthcare (surgical technology, detection of new, previously undiagnosed diseases, identification of new modes of treatment), optimization of transportation, factories, and decision-making improvement, among various other applications.

For this reason, we advocate for a proper use where technological developments always benefit humanity. It is crucial to ensure that the implementation of artificial intelligence is carried out contributing positively to the well-being and progress of humanity.

The incorporation of technology and, particularly, artificial intelligence in human relations calls for a reconsideration of Ecobioethics, leading to the development of a specialized branch focused on the interaction with this type of technology – Ecotechnobioethics. This will enable the protection and safeguarding of human subjectivity against the advancement of AI.

It is interesting to clarify that technology does not happen to us; it is a human creation motivated by needs that we impose. In this way, we are not only in relation to technology but perceive it as something inherent to the unfolding of our own subjectivity. Thus, in the face of technological advances and especially AI, the impact they generate on our psyche compels us to study in-depth the consequences they can have on subjectivity.

6. Conclusions

The development of a new area within Ecobioethics proposes a specialized dimension of analysis regarding Ecobioethical considerations in the interaction between artificial intelligence (AI) and humans. Ecotechnobioethics emerges as a guiding light for the development of AI, with the aim, between others, of preserving mental health, fundamental human rights, and considering appropriate use in its developments. Specific cases, such as Digital Necromancy, underscore the urgent need for the development of Ecotechnobioethics to address these possibly manifestations in a specialized manner, from the grieving process to post-mortem intellectual property. Positioned at the intersection of technology and human subjectivity, Ecotechnobioethics not only provides steps to address the challenges presented but also emphasizes the urgent need to establish norms and regulations that safeguard human integrity in this new and complex technological terrain.

References

1. Colmenares Suarez, L (2018) De la bioética a la ecobioetica: una breve referencia a su desarrollo histórico. Rev. Centro medico de Caracas. Vol.57, No. 146.
2. Benyakar, M (2016) Un trayecto de vida. Fundaciónkonex.org https://www.fundacionkonex.org/b4830-moty-benyakar
3. Benyakar, M et all (2005) Lo traumático, clinica y paradoja tomo 1-1ª Ed- Buenos Aires: Biblos.
4. Benyakar, M et all (2006) Lo traumático, clinica y paradoja tomo 2-1ª Ed- Buenos Aires: Biblos.
5. Benyakar, M (2006) Lo disruptivo, amenazas individuales y colectivas: el psquismo ante guerras, terrorismos y catástrofes sociales – 2ª ed.- Buenos Aires: Biblos
6. Corominas, J (1987) Diccionario Etimológico de la Lengua Castellana. Buenos Aires. Editorial Gredos, S.A.
7. Lacan, J (2008) El Seminario, libro 11: Los cuatro conceptos fundamentales del psicoanálisis, Paidós, Buenos Aires, p. 139.
8. Benyakar, M (2024) Impactos Disruptivos por entorno en la Era Contemporánea Etérea. (in press)
9. Brooker, D et all. (2023) Nigromancia digital: cómo la inteligencia artificial cambia nuestra relación con los muertos. BBC.com url: https://www.bbc.com/mundo/articles/crg1d98lv9wo
10. Freud, S (1917) “Duelo y melancolía”, Obras Completas, Tomo XIV, Amorrortu Editores, Buenos Aires,
11. Brundage, M., Amodei, D., Russell, C., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

12. Wallach, W., Allen, C., y Smit, J. (2012). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.
13. Beauchamp, T. L., y Childress, J. F. (2013). Principles of Biomedical Ethics (8ª ed.). Oxford University Press.
14. Floridi, L. (2016). The Ethics of Artificial Intelligence. Oxford University Press.
15. Brundage, M., Amodei, D., Russell, C., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
16. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
17. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.
18. Calo, R., y McQuade, S. (2018). The right to explanation: Lessons from machine learning and artificial intelligence. Washington Law Review, 93(5), 1785-1832.
19. Gaut, B (2013) Art and Ethics. 3rd edition. Routledge ISBN 9780203813034

4. Safety and public interest

5. AI and Surgery

6. Education and Training

Ethical Frontiers in AI-Enabled Medical Education: Considerations on the Role of Large Language Models in Medical Learning

Gerson Hiroshi Yoshinari Júnior MD, PhD,
Faculty of Medicine of Itajubá, Itajubá, Minas Gerais, Brazil.
Email:gerson.junior@fmit.edu.br
ORCID: 0000-0001-5837-5743
Luciano Magalhães Vitorino RN, PhD, RN.
Faculty of Medicine of Itajubá, Itajubá, Minas Gerais, Brazil.
Email: lucianoenf@yahoo.com.br.
ORCID: 0000-0003-1023-8488

Introduction to AI in Medical and Health Education

In the rapidly evolving landscape of education, Artificial Intelligence (AI) stands out as a transformative force, reshaping traditional learning into dynamic and interactive experiences. The integration of AI within education, especially in medical and health disciplines, marks a shift towards more personalized and adaptive learning environments (1, 2). AI's role extends from streamlining administrative tasks to offering bespoke learning journeys, highlighting its potential to revolutionize education by making it more attuned to individual needs and the demands of modern healthcare practices (3, 4).

AI's impact on medical education is particularly significant, with its applications ranging from virtual patient simulations to adaptive learning algorithms(5). Such innovations enhance the educational experience, preparing students for the nuanced realities of healthcare. Virtual simulations allow for the practice of diagnostic and decision-making skills within a safe, controlled environment, enhancing the comprehension of medical principles and patient care (6). This not only improves educational outcomes but also bridges the gap between theoretical knowledge and practical application.

Furthermore, AI supports educators by automating routine tasks, such as grading and attendance tracking, thereby enabling a focus on personalized interaction and teaching. Advanced analytics powered by AI assist in identifying students needing extra

support, facilitating timely intervention and customization of help. AI's capability to keep educational content current, by integrating the latest medical research into learning materials, ensures that students are learning from the most recent and relevant information.

For students, AI provides a learning experience tailored to their individual pace and style, through intelligent tutoring systems that offer personalized feedback and guidance. Virtual simulations present opportunities for hands-on practice without the risks associated with real-life patient care. AI also fosters a collaborative learning environment by recommending study groups or partners, enhancing the sense of community within the educational setting (7).

The success of AI in education is contingent on the availability of rich medical datasets, which fuel the realism of simulations and the effectiveness of personalized learning paths. This reliance on data highlights critical ethical considerations, including privacy, consent, and data security, setting a foundation for a deeper exploration of the ethical dimensions of AI in medical education (8). As we navigate these considerations, the importance of ethical data use becomes a central theme, leading to a broader discussion on ensuring AI's responsible and beneficial integration into medical and health education.

Data Ethics, Bias, and Transparency in Medical AI Education

Large Language Models (LLMs) like ChatGPT epitomize the advancements in AI, offering capabilities to parse and analyze extensive datasets to mimic human language nuances. This process, reliant on vast amounts of diverse text data, is designed to capture the complexity of human communication across various contexts, enabling LLMs to produce responses that closely resemble human discourse. However, this dependency on large datasets introduces significant ethical considerations, particularly the potential for inherent biases. These biases may stem from unrepresentative training data, skewing AI outputs and reflecting existing disparities within the model's decision-making processes.

The phenomenon of "hallucinations," where LLMs generate information not directly derived from their training data, illustrates another layer of complexity, leading to potential inaccuracies. This, combined with the models' "black box" nature—characterized by their intricate and opaque processing—poses challenges in diagnosing and correcting biases or inaccuracies(9). Such opacity underscores the urgent need for improved transparency and interpretability within AI technologies, especially in educational settings where the stakes for accuracy and fairness are notably high.

In medical and health education, the ethical landscape surrounding the use of LLMs and other AI technologies is particularly fraught with challenges around data privacy, confidentiality, and algorithmic bias(1, 6). The necessity for comprehensive medical datasets to fuel AI-driven educational tools underscores significant ethical dilemmas, especially regarding the potential misuse of sensitive information. Moreover, biases in AI algorithms can exacerbate societal and healthcare inequities, impacting educational content and students' clinical judgment (10, 11). To address these issues, there's a critical need for integrating discussions on data ethics, privacy, and fairness into the medical curriculum, preparing students to confront and navigate these ethical complexities in their future practice.

Achieving transparency and explainability in AI within medical education is paramount, ensuring systems are not only accurate but also understandable to users. The challenge lies in the inherent complexity of AI systems, particularly those based on deep learning, which complicates efforts to make AI's decision-making processes transparent and interpretable. Addressing these challenges requires a concerted effort from all stakeholders involved in AI's educational integration, advocating for the development of AI systems that are both powerful and accountable(11). By prioritizing transparency and ethical considerations, such as data ethics and bias mitigation, we can ensure that AI technologies are effectively and responsibly harnessed to enrich medical education, ultimately leading to a healthcare workforce that is both technologically proficient and ethically informed.

Impact on the Educator and Learner Relationship

The advent of AI in medical and health education heralds a transformative shift in the dynamics between educators and learners, reshaping the traditional roles and interactions that have long characterized the educational landscape(11). AI's capability to personalize learning, automate administrative tasks, and provide instant feedback has the potential to greatly enhance the efficiency and effectiveness of education. However, this technological intervention also raises profound questions about the future role of human educators and the nature of the student-teacher relationship. As AI tools become more integrated into educational settings, there is a growing potential for these technologies to supplement, and in some cases, substitute for human instruction, prompting a reevaluation of what it means to teach and learn in the digital age (10, 12).

The potential for AI to supplement human instruction offers exciting possibilities for enhancing educational quality and accessibility. AI-driven platforms can provide personalized learning experiences, adapt to individual student needs, and offer insights into student progress that can inform teaching strategies. This level of customization and support can free educators from routine tasks, allowing them to focus more on the nuanced aspects of teaching that require empathy, moral judgment, and critical thinking—qualities that AI cannot replicate (13). However, the prospect of AI substituting human educators in certain contexts raises ethical concerns regarding the depersonalization of education, the loss of direct human mentorship, and the potential erosion of ethical standards that human educators uphold.

The implications of these changes for educational quality and ethical standards are profound. While AI can enhance the learning experience by providing data-driven insights and freeing educators to engage in more meaningful interactions with students, the essence of education—particularly in fields as intrinsically human as medicine and health—lies in the relationship between educator and learner. This relationship is built on trust, empathy, and a deep understanding of ethical principles, aspects that AI cannot fully replicate. Therefore, as we navigate the integration of AI into educational contexts, it is crucial to maintain a balance that leverages the strengths of AI to enhance education while preserving the irreplaceable value of human interaction and ethical guidance in the learning process.

AI Accountability and Regulation in Education

The advent of AI in education has ushered in a transformative era, marked by innovations that promise to redefine teaching and learning paradigms. However, this digital renaissance is accompanied by a complex web of ethical, legal, and regulatory challenges that necessitate a nuanced understanding of the evolving legislative landscape, particularly within the European Union (EU)(14).

The EU's pioneering legislative framework, the AI Act, emerges as a cornerstone in the global discourse on AI governance, setting a precedent for a risk-based regulatory approach. This landmark Act categorizes AI systems according to the risk they pose to society, ranging from high-risk applications, which are subject to stringent compliance requirements, to low or minimal risk applications, where regulatory obligations are significantly lighter(15).

High-Risk Applications in Education

Within the educational sphere, AI systems classified as high-risk under the AI Act are subject to rigorous scrutiny. These systems, which might include AI-driven examination tools or personalized learning platforms, are pivotal in shaping educational outcomes and thus bear a substantial impact on students' futures. The Act mandates these high-risk applications to adhere to strict compliance measures, including transparency, data governance, and accuracy standards, to ensure their ethical deployment and operational integrity.

The obligation for high-risk AI systems to undergo thorough assessment processes before deployment underscores the EU's commitment to safeguarding educational equity and upholding students' rights. It reflects a conscientious effort to balance innovation with accountability, ensuring that AI applications in education enhance rather than undermine the learning experience(15).

Low or Minimal Risk: A Space for Innovation

Conversely, AI applications deemed to pose low or minimal risk are afforded a degree of regulatory latitude under the AI Act. This categorization encompasses a broad spectrum of educational tools and platforms that, while transformative, are considered less likely to have a significant adverse impact on students or the educational process. The Act's nuanced approach in this regard encourages innovation and experimentation within the educational technology sector, fostering an environment where AI can be leveraged to enhance learning outcomes without the burden of onerous regulation(15).

Unacceptable Risk: The Red Lines

At the opposite end of the spectrum, the AI Act identifies certain uses of AI that are considered to pose an unacceptable risk and are, therefore, prohibited. In the context of education, this could relate to AI systems that manipulate student behaviors or exploit vulnerabilities, representing a clear violation of ethical principles and students' rights. By delineating these red lines, the AI Act ensures that the integration of AI into education respects the foundational values of trust, dignity, and respect for individual autonomy(15).

Ethical Frameworks for AI in Education

As Artificial Intelligence (AI) becomes increasingly embedded in medical and health education, it presents a dual-edged sword—offering remarkable tools for enhancing learning and decision-making while raising significant ethical dilemmas. The deployment of AI in educational settings necessitates a conscientious approach, underpinned by ethical frameworks that ensure technologies are used to benefit students ethically and equitably. At the heart of these considerations are the principles of autonomy, justice, beneficence, and non-maleficence, each playing a crucial role in guiding the responsible integration of AI into educational practices(16, 17).

The principle of autonomy emphasizes the necessity of empowering users—both students and educators—by ensuring they have the knowledge and freedom to engage with AI technologies on their terms. This entails transparent information about how AI systems make decisions, the data they use, and the potential biases they may introduce.

Such transparency is crucial for maintaining trust and agency in educational environments, where the goal is to enhance, not undermine, the learning experience. Justice, in this context, extends beyond mere access to AI tools; it encompasses the fair distribution of their benefits and the mitigation of risks across diverse student populations. Ensuring that AI-driven educational tools do not widen the gap but rather foster inclusivity and equality is paramount(16, 17).

Beneficence and non-maleficence serve as the ethical pillars that demand AI applications in education not only do good but also avoid harm. Beneficence calls for the development and application of AI that genuinely enhances educational outcomes, providing students with engaging, effective, and personalized learning experiences. Conversely, non-maleficence requires vigilance against the potential negative impacts of AI, such as reinforcing biases, compromising privacy, or diminishing the human interaction that is critical to the educational experience. By adhering to these principles, stakeholders can navigate the complex ethical landscape of AI in education, ensuring that technological advancements serve to uplift and equalize, rather than to divide or diminish(16, 17).

Future perspectives

As we peer into the future of work in health professions, it becomes evident that Artificial Intelligence (AI) education tools are not just supplementary technologies but fundamental elements reshaping the landscape of healthcare education and practice. The integration of AI into healthcare education heralds a transformative shift, where the skills and competencies required for future healthcare professionals will be markedly different from those of today. Embracing AI tools in education equips students with a sophisticated understanding of data analysis, machine learning, and digital ethics, preparing them for a healthcare environment where AI assists in diagnostics, treatment planning, and patient care. This preparation not only enhances their clinical skills but also cultivates a mindset that is critical, adaptive, and ethically attuned to the nuances of technology-driven healthcare.

However, this evolution is not without its ethical challenges. The increasing reliance on AI in healthcare education necessitates a careful balance between leveraging technology for educational advancement and preserving the humanistic values central to healthcare. The potential for AI to inadvertently perpetuate biases, infringe on privacy, or devalue the importance of empathy and interpersonal skills in healthcare underscores the need for ethical vigilance. Educators, developers, and policymakers must collaborate to ensure that AI tools are designed and used in ways that uphold ethical standards and foster an environment where technology enhances, rather than detracts from, the human aspects of healthcare.

Looking ahead, the convergence of AI and healthcare education presents both profound opportunities and complex ethical dilemmas. As we navigate this evolving landscape, it is crucial to remain committed to principles of equity, transparency, and accountability. Ensuring that AI technologies are accessible to all students, transparent in their operations and decision-making processes, and accountable for the outcomes they produce is essential. This commitment not only safeguards against ethical pitfalls but also ensures that AI serves as a catalyst for enriching medical and health education, ultimately leading to more competent, compassionate, and technologically proficient healthcare professionals.

Conclusion

In conclusion, the integration of AI into medical and health education demands a forward-looking approach that embraces the potential of technology while steadfastly adhering to ethical principles. As we venture into this new era, recommendations for educators, developers, policymakers, and researchers center on fostering an educational ecosystem where AI is used responsibly. This entails rigorous ethical oversight, ongoing professional development for educators in the ethical use of AI, and the development of curricula that balance technological proficiency with the enduring values of care and compassion. By embracing these challenges and opportunities, we can ensure that the future of healthcare is not only technologically advanced but also ethically grounded and human-centric.

References

1. Masters K. Artificial intelligence in medical education. Med Teach. 2019;41(9):976-80.
2. Vitorino LM, Yoshinari Júnior GH. Artificial intelligence as an ally in Brazilian nursing: challenges, opportunities and professional responsibility. Rev Bras Enferm. 2023;76(3):e760301.
3. Parycek P, Schmid V, Novak A-S. Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions. Journal of the Knowledge Economy. 2023:1-26.
4. Vitorino LM, Júnior GHY. ChatGPT and the teaching of contemporary nursing: And now professor? J Clin Nurs. 2023;32(21-22):7921-2.
5. Yoshinari Júnior GH, Vitorino LM. How may ChatGPT impact medical teaching? Rev Assoc Med Bras (1992). 2023;69(4):e20230282.
6. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in Medical Education: Global situation, effects and challenges. Education and Information Technologies. 2023:1-23.
7. Seo K, Tang J, Roll I, Fels S, Yoon D. The impact of artificial intelligence on learner–instructor interaction in online learning. International Journal of Educational Technology in Higher Education. 2021;18(1):54.
8. Nguyen A, Ngo HN, Hong Y, Dang B, Nguyen B-PT. Ethical principles for artificial intelligence in education. Education and Information Technologies. 2023;28(4):4221-41.
9. Bruno A, Mazzeo PL, Chetouani A, Tliba M, Kerkouri MA. Insights into Classifying and Mitigating LLMs' Hallucinations. arXiv preprint arXiv:231108117. 2023.
10. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. 2021;5(6):493-7.
11. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. The Lancet Digital Health. 2020;2(3):e111-e2.
12. Lakhani HV, Pillai SS, Zehra M, Sharma I, Sodhi K. Systematic Review of Clinical Insights into Novel Coronavirus (CoVID-19) Pandemic: Persisting Challenges in U.S. Rural Population. Int J Environ Res Public Health. 2020;17(12).
13. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nature Medicine. 2023;29(11):2929-38.
14. MADIEGA TA. EU guidelines on ethics in artificial intelligence: Context and implementation. 2019.
15. Thelisson E, Verma H. Conformity assessment under the EU AI act general approach. AI and Ethics. 2024.
16. Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines. 2020;30(1):99-120.
17. Varkey B. Principles of Clinical Ethics and Their Application to Practice. Med Princ Pract. 2021;30(1):17-28.

Transforming Public Healthcare and Education Using AI-powered Mixed Reality Technology

Predrag Stevanović[1,2], Lazar Davidović[1,3]

Prof. Predrag Stevanović, MD, PhD
[1]Medical Faculty, University of Belgrade, Serbia
[2]Clinical Hospital Center "Dr. Dragiša Mišović – Dedinje", Belgrade Serbia

Prof. Lazar Davidović, MD, PhD
[1]Dean of the Faculty of Medicine, University of Belgrade, Serbia
[3]Clinic for Vascular Surgery, University Clinical Centre of Serbia, Belgrade, Serbia

Artificial intelligence (AI) can be broadly described as the capacity of a computer or device to analyze extensive and intricate data, uncover insights, detect potential risks and opportunities, and facilitate better decision-making. In the rapidly evolving field of AI, prominent techniques employed in healthcare comprise machine learning, natural language processing, and the integration of AI with clinical decision support systems, often accomplished through the creation of user-friendly graphical interfaces.[1]

Machine learning is a prevalent branch of artificial intelligence (AI) that involves using statistical techniques to train models on data.[1] The computer "learns" to comprehend the data by analyzing training datasets as examples. Looking ahead, AI holds promising future roles in clinical decision support systems that enhance patient safety. AI can contribute to perioperative patient safety by enabling earlier detection of clinical deterioration and offering clinical decision support for optimal management of intraoperative physiological changes. An illustration of this can be seen in the Hypotension Prediction Index, which is currently being utilized in everyday anesthesia practice.[2] By employing machine learning techniques and gathering data from the patient's vital signs monitors, the program can forecast the onset of hypotension, enabling the anaesthesiologist to intervene and take necessary measures proactively.

On the other hand, mixed reality (MR), also known as hybrid reality, represents an emerging technology that amalgamates elements from both virtual reality (VR) and augmented reality (AR) to create immersive and interactive experiences. Mixed/augmented reality represents a combination of physical and digital worlds in which users can interact with digital objects (while keeping their presence in the physical, real world). Mixed reality holds considerable promise in modern medicine and is increasingly being explored for potential applications.
Public healthcare in Serbia is transformed through the innovative use of AI-powered MR technology, which increases healthcare efficiency and quality, minimizes risks and efforts, and optimizes procedures. One such device is Microsoft's HoloLens2 (HL2), which enables the use of artificial intelligence (AI) and mixed/augmented reality (MR/AR) in everyday practice. Through remote collaboration, doctors have the same insight into a patient's condition without physical presence, enabling joint real-time inputs and medical interventions with experts from anywhere (Figure 1) and remote education for medical students.



Figure 1. Long-distance professional collaboration using AI-powered MR technology

Presented below are vital insights into the role of AI-powered mixed reality in contemporary healthcare:

- Surgical Planning and Training: Mixed reality facilitates the visualization and manipulation of three-dimensional anatomical models and medical imaging data in real-time, empowering surgeons with an intuitive and immersive platform for surgical planning. This technology enables the simulation of procedures, the practice of techniques, and the refinement of skills within a risk-free mixed/virtual environment. As a result, surgical precision and efficiency are enhanced, ultimately improving patient outcomes.[3]
- Real-time Intraoperative Assistance: By superimposing pertinent patient data, such as CT scans or MRI images, onto the surgical field, mixed reality offers surgeons real-time guidance and support during surgical procedures. This integration enables immediate access to critical information without diverting attention from the task. Consequently, accuracy is improved, errors are reduced, and decision-making during complex surgeries is enhanced.[4]
- Medical Education and Simulation: Mixed reality presents a powerful medical education and training tool. It allows medical students, residents, and healthcare professionals to visualize intricate medical concepts, engage in clinical scenarios, and simulate patient interactions within a highly realistic virtual environment. This immersive learning experience accelerates learning, enhances knowledge retention, and bridges the theoretical understanding and practical application gap.[5]
- Patient Education and Rehabilitation: Mixed reality facilitates patient education by delivering interactive and personalized visualizations of medical conditions, treatment options, and procedural explanations. This technology empowers patients to comprehend their conditions more effectively, make informed decisions regarding their healthcare, and actively engage in their medical journey. Additionally, mixed reality can be employed in rehabilitation settings to design engaging and motivating exercises for physical therapy and cognitive rehabilitation.[6]
- Remote Consultations and Telemedicine: The incorporation of mixed reality into remote consultations and telemedicine has the potential to enrich the virtual healthcare experience. By integrating 3D models, medical imaging, and real-time data visualization, healthcare providers can remotely assess and diagnose patients in a more immersive and accurate manner. This technology transcends geographical barriers, improves access to specialized care, and enhances the efficiency of remote healthcare delivery.[7]
- Pain Management and Distraction Techniques: Mixed reality has been investigated to manage pain and provide distraction during medical procedures. By immersing patients in virtual

environments or presenting them with interactive experiences, mixed reality can divert attention from pain stimuli, alleviate anxiety, and enhance patient comfort. This application has demonstrated potential across diverse settings, including pediatric care, dental procedures, and wound care.[8]

In contrast to virtual reality, which allows users to interact with entirely artificial environments, augmented (mixed) reality generates 3D computer objects onto real-world surfaces, thereby providing a combined stereoscopic visualization. While observing the physical environment, users can manipulate digital content through holograms generated by the device.[9] All the benefits of mixed reality can be harnessed through various innovative technological solutions. One of the devices is Microsoft's HoloLens 2 (HL-2), which enables AI and MR in everyday practice.

Introducing the Microsoft's HoloLens 2

The HL-2 is a head-mounted display unit that establishes a connection with a remote cloud infrastructure for image reconstruction and storage of audio-video data (Figure 2). The frontal section of the device houses a collection of sensors along with their corresponding hardware components, such as processors, cameras, and projection lenses. The device settings can be tailored to accommodate the visual characteristics of the individual user.[10] Additionally, the device incorporates two 3D audio speakers positioned near the user's ears, enabling the simultaneous perception of sounds from both the physical environment and the virtual reality environment.[11] The device features transparent holographic lenses that provide a clear view of the virtual content overlaid in the real world.



Figure 2. Hololens 2 headset

The HL2 device incorporates AI-powered machine learning algorithms that enable it to accurately map and analyze its surrounding space, thereby acquiring an understanding of its unique characteristics. Furthermore, it can retain this learned memory of the spatial environment, ensuring its persistence even when the device is powered off after activating the machine learning mode. This feature allows the device to recognize and recall the entire room consistently, even during repeated entries.
The hands-free user interface of the HL2 device is non-intrusive; the user can easily see data outside the main view field, which is important for primary work. The HL2 device's AI recognizes hand gestures that the user makes while controlling the device; this is a key component of hands-free operation.

In addition to hand gestures, the device AI also observes the user's eye movement and detects when the user focuses on an option offered by the user interface (UI); extra options can be selected by eye-blinking. Furthermore, the HL2 is also equipped with speech-recognition technology as another set of AI-powered tools that can be used for UI selection and environment control. More importantly, the speech-recognition capability is used for dictating patient notes; these notes can be stored, shared, or even preserved in the same spatial environment for all other staff members who use the device in the same workspace. The HL2 device's extensive capabilities contributed to increased work efficiency.
The display resolution is 2k with a 3:2 aspect ratio, powered by an LED light engine. Eye-based rendering is employed, based on the position of the user's eyes, to optimize the display for a three-dimensional viewing experience. The device incorporates tracking technology to monitor the movement of the user's head and eyes. The Azure Kinect sensor facilitates depth perception, while motion sensing is achieved through an accelerometer, gyroscope, and magnetometer. Additionally, the device is equipped with an 8MP camera capable of capturing still images and recording 1080p30 video.[10,11]
Extensive research examined the utilization of augmented reality within diverse medical domains, encompassing surgical settings, emergency units, and medical education. The characteristics of this particular device were examined in various clinical situations, and a brief report summarizing the findings is presented.

The HoloLens 2 usage during the COVID-19 pandemic

The COVID-19 pandemic has significantly strained even the most advanced healthcare systems, posing numerous challenges to traditional hospital treatment.[12] Safeguarding the well-being and safety of healthcare personnel was paramount for upholding the quality of patient care and sustaining the capacity to manage escalated demands. The rapid surge of COVID-19 patients within a condensed timeframe resulted in an overwhelming engagement of healthcare professionals, intensifying chronic stress and work-related fatigue.[13] Prolonged work shifts, the risk of infection transmission, and the protracted duration of the pandemic necessitated reducing staff exposure to highly contagious SARS-CoV-2 environments. Including a substantial number of less experienced clinicians in COVID zones introduced the risk of communication errors and disruptions in the continuity of treatment. The substantial consumption of personal protective equipment (PPE) during the pandemic engendered an incessant risk of shortages, prompting the urgent need for novel methodologies to optimize its utilization.[14] Certain studies documented the implementation of HL2 to minimize healthcare professionals' exposure to aerosol-generating procedures in nephrology wards amidst the COVID-19 pandemic.[15,16]
In the Clinical Hospital Center "Dr Dragiša Mišović – Dedinje" , which is part of the Medical Faculty of the University of Belgrade, a study was conducted with the primary objective to examine the potential reduction in doctors' exposure time in the COVID-19 intensive care unit (ICU) red zone by utilizing an HL-2 device. Additionally, we sought to evaluate the usability and acceptability of this innovative technology. The duration of doctors' exposure to the COVID-19 infectious agent was measured by recording the total time spent in the COVID-19 ICU during the morning shift and the frequency of entries into the COVID-19 ICU. These variables were documented for each doctor over 30 days, including 15 days without the HL-2 device and 15 days with its use. During the HL-2 device phase, one doctor from the team would enter the COVID-19 ICU red zone while utilizing the device.
Upon completing the study, we calculated the average daily time spent by each doctor in the COVID-19 ICU. We also conducted a survey (enclosed) to gather anaesthesiologists' feedback on the usability, acceptability, and satisfaction with the HL-2 technology.
Throughout the study period, the morning ward rounds were initiated by one doctor from the morning shift. At the same time, the rest of the team participated remotely from the green zone (outside the COVID-19 ICU) via the Microsoft Teams platform. The doctor inside the COVID-19 ICU and the team

members outside the red zone had a direct view of patients, monitoring devices, and ventilator or life support device parameters (Figures 3 and 4).

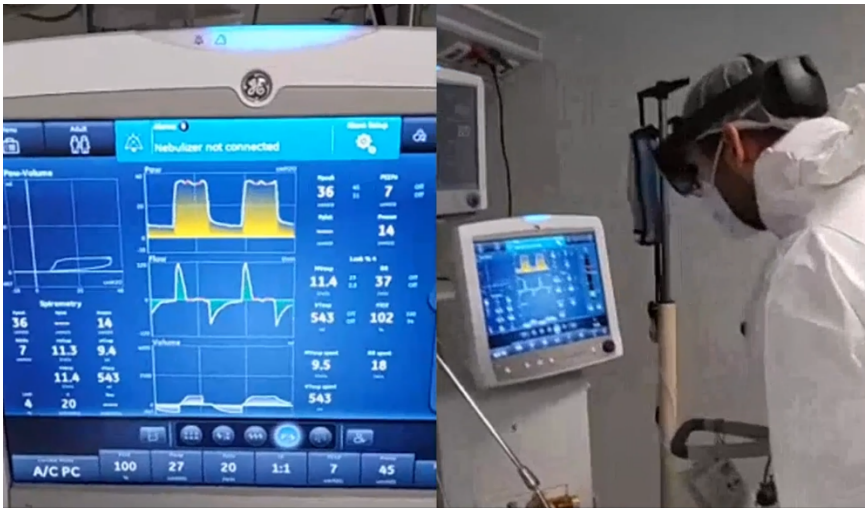
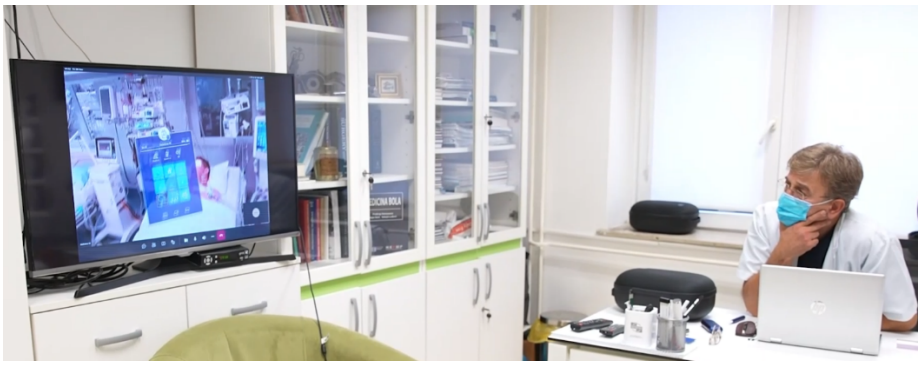


Figure 3. Long-distance professional collaboration during the COVID-19 pandemic (a)


Figure 4. Long-distance professional collaboration during the COVID-19 pandemic (b)

Furthermore, the MR technology of the device allowed all participants to visualize laboratory results, radiological diagnostic findings, and notes containing personal observations from the previous shift physicians as holograms displayed within the patient's visual proximity.

The study comprised 21 anesthesiologists. The average time spent in the red zone per doctor showed a significant decrease ($p<0.001$) of 74±52 (95% CI 50-97) minutes, representing an average reduction of 20.5%, from 361±45 (95% CI 341-381) minutes in regular mode to 287±33 (95% CI 272-302) minutes in HoloLens mode. Overall, 85.7% of the participating physicians reported being satisfied with using the HL2 device in treating critically ill COVID-19 patients.

HoloLens 2 as a telemedicine tool in endovascular procedures

The prevalence of vascular diseases increases as life expectancy rises, leading to complex cases involving elderly patients with multiple comorbidities.[17] Vascular surgery has become an independent surgical specialization characterized by its dynamic nature and reliance on technology, offering various treatment options.

Traditional medicine offers surgical repair as the only preventive solution for abdominal aneurysm rupture, making abdominal aneurysm surgery a complex procedure. In recent decades less invasive vascular and endovascular techniques have emerged as acceptable treatment options. Endovascular techniques, performed through groin or axillary access, offer the advantage of avoiding opening the abdominal or thoracic cavities. Thorough planning is essential for procedures performed remotely, such as implanting a stent graft controlled from the groin, while imaging technology is critical in planning and execution. Image fusion techniques combining preoperative and intraoperative images are frequently used to enhance procedure accuracy, reduce radiation exposure, and minimize contrast medium usage. Accurate image processing and interpretation are crucial for preoperative planning and multidisciplinary team assessments.

At the Clinic for Vascular Surgery, Clinical Centre of Serbia in Belgrade, as a base of the Medical Faculty of the University of Belgrade, under the supervision of Prof. Lazar Davidović, a challenging case was presented and discussed in a multidisciplinary meeting across different countries, aided by MR for data sharing. The case was presented using HL-2 technology to an expert colleague from Malta, Prof. Kevin Cassar, who contributed with his experience in treating similar complex patients. The reported case involved a patient with a juxtarenal abdominal aortic aneurysm, cardio and respiratory comorbidities and a hostile abdomen due to previous surgeries. MR allowed for a detailed presentation of all aspects of the problem, facilitating professional and realistic discussions and decision-making (Figure 5).



Figure 5. Long-distance professional collaboration between Serbia and Malta (a)

During the surgical procedure, MR continued to share information, show intraoperative findings, and discuss the final result. After consultations among vascular surgeons, interventional radiologists, an anesthesiologist in the operating room, and the remote expert professor, it was decided to treat the patient using endovascular techniques. The procedure involved using a “chimney” parallel graft for one renal artery and the implantation of a stent graft in the infrarenal aorta, with incisions made in the bilateral groin and right brachial area. The operator in the operating room, Prof. Igor Končar wore HoloLens glasses to assist with images, enabling interactive and immediate consultation by comparing previous and current images without the need for image fusion technology (Figure 6). Wearing the glasses did not disrupt the operator’s work, as the digital reality was limited to their peripheral vision. The patient was discharged home the following day.



Figure 6. Long-distance professional collaboration between Serbia and Malta (b)

The Role of HoloLens 2 in remote collaboration for pain treatment

Postlaminectomy syndrome, also known as Failed Back Surgery Syndrome (FBSS), is a relatively common cause of pain following spinal surgery.[18] One of the treatment methods for this painful syndrome is epidural lysis or epidurolysis.[18] Clinical Hospital Centre "Dr Dragiša Mišović – Dedinje", part of the Medical Faculty of the University of Belgrade, epidurolysis is performed using the FORA B catheter (Seawon Meditech). In the case of a young female patient, she experienced severe lower back pain with radicular presentation after multiple surgeries in the L4-S1 segment, including stabilization. Postoperative sequelae with fibrous tissue in the epidural space were observed on MRI. Anticipating difficulties in guiding the catheter toward the fibrotic region, we planned to utilize HoloLens technology for communication with Prof. Andreas Veihelmann from Germany (Figure 7). Indeed, the catheter navigation from the sacral hiatus to the targeted region at the L5-S1 transition was challenging, and Prof. Veihelmann provided remote assistance.



Figure 7. Long-distance professional collaboration between Serbia and Germany

The AI program can render imported radiological findings within HL-2, generating 3D models that can be projected onto the patient's body as holograms. These holographic representations enable a more precise orientation towards the target pathology during the execution of interventions. We utilized HL2 to present a 3D reconstruction of the spinal column and canal and MRI sections (Figures 8 and 9). Dr.

Veihelmann could follow our work and the fluoroscopic image that tracked the catheter's movement. Through discussion and advice from our colleague in Germany, Prof. Predrag Stevanović successfully performed the intervention.

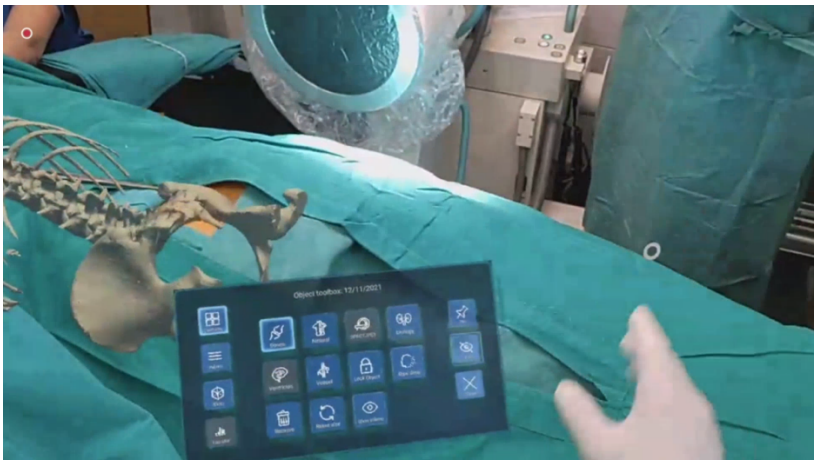


Figure 8. Artificial intelligence segmentation tools

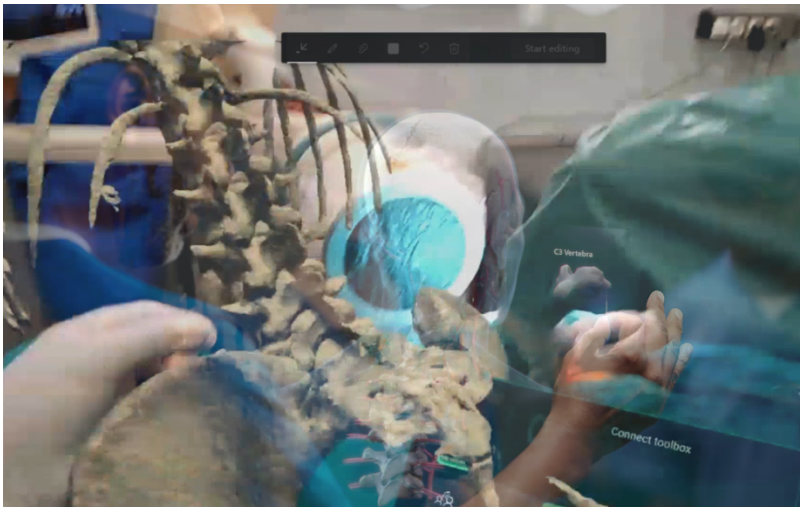


Figure 9. Visualization of real-world medical data

The utilization of Mixed Reality (MR) in medical education

MR has revolutionized the field of simulation-based training. MR offers a unique and immersive platform that enables medical students and healthcare professionals to engage in realistic and interactive learning experiences.[19] Simulations using MR technology replicate clinical scenarios, procedures, and anatomical structures, allowing learners to practice and refine their skills in a safe and controlled environment.[20]

One of the significant advantages of MR simulation in medical education is its ability to bridge the gap between theoretical knowledge and practical application. MR simulations also provide opportunities for learners to practice clinical decision-making and critical thinking in a risk-free environment. They can simulate patient encounters, diagnose conditions, and make treatment decisions based on realistic scenarios.[20]

At the University of Belgrade, Faculty of Medicine in Serbia, students can engage in a simulation center utilizing HL2 and MR technology, allowing them to acquaint themselves with diverse clinical scenarios

and the execution of numerous procedures. While students work with their instructor in the simulation center, a fellow anesthesiologist employs HL2 in a remote ICU with actual patients (Figure 10). Real-time video footage from the intensive care unit enables students to observe the dynamic adjustment of mechanical ventilator parameters tailored to each patient's specific needs and the placement of central venous catheters and arterial lines, supplemented by expert explanations from the intensivist.

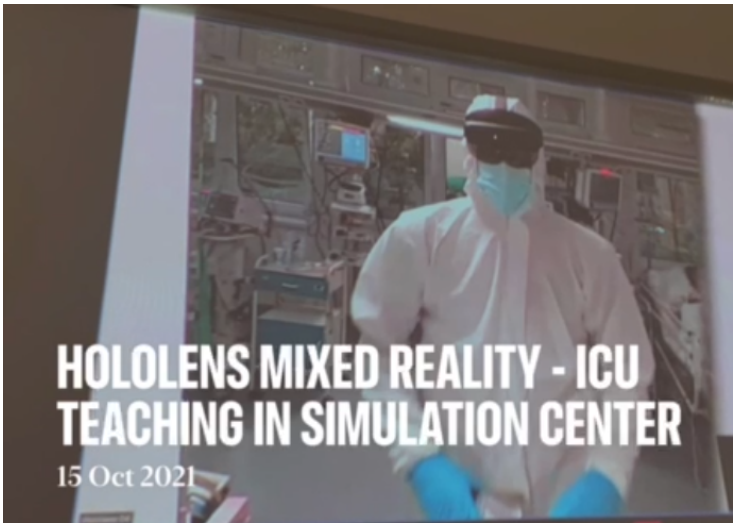


Figure 10. HoloLens 2 mixed reality – Intensive care unit teaching in the simulation center

Furthermore, HL-2 facilitates students' immersive experience by simulating varying clinical conditions at different stages, fostering discussion and collective decision-making with mentors regarding subsequent diagnostic and therapeutic interventions. A noteworthy instance of this immersive learning occurred during the 2022 Student Congress, where students had the opportunity to utilize HL2, witnessing the progression of a patient's deteriorating state due to chronic obstructive pulmonary disease, an epileptic seizure, comatose status, and Parkinson's disease. Each patient was represented as a virtual 3D animation, allowing students to manipulate the imagery while concurrently monitoring the patients' vital parameters on adjacent displays (Figure 11). The utilization of HL-2 and students' exposure to clinical practice has constituted an indelible and invaluable experience for many.


Figure 11. Medical students learning from virtual cases

Using MR in medical education through simulation offers a powerful and transformative learning tool. It provides an immersive, interactive, and safe environment for learners to acquire and refine clinical skills, enhance decision-making, and prepare for real-world healthcare scenarios. As technology advances, MR simulations hold great promise in shaping the future of medical education and training (Figure 12).



Figure 12. Hologram of the patient positioned in the audience. Demonstration of simulation scenarios to a large number of observers

While mixed reality presents numerous benefits, its adoption within modern medicine necessitates addressing technical challenges, ensuring data privacy and security, and substantiating its efficacy through rigorous research and clinical trials. Collaboration among technology developers, healthcare professionals, and regulatory bodies is paramount in fully harnessing the potential of mixed reality to advance healthcare delivery and enhance patient outcomes.

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-40.
2. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TW. Ability of an arterial waveform analysis–derived hypotension prediction index to predict future hypotensive events in surgical patients. Anesth Analg. 2020;130(2):352-9.
3. Lozano-Quilis JA, Pizarro-Romero P, Azorin-Lopez J, et al. Mixed reality in surgical planning and training. IEEE Access. 2019;7:153353-153368. doi: 10.1109/ACCESS.2019.2949611
4. Muensterer OJ, Lacher M, Zoeller C, Bronstein M, Kübler J. Microsoft HoloLens in mixed reality-assisted surgery: An experience on feasibility and accuracy of navigational information in pediatric surgery. Eur J Pediatr Surg. 2017;27(3):384-391. doi: 10.1055/s-0037-1605392
5. Kusaslan Avci D, Yilmaz R, Pehlivan Z, et al. Mixed reality for medical education and training. In: 2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE; 2019:1-4. doi: 10.1109/SIU.2019.8806314
6. Cheung B, Lo A, Law B, So H. Mixed reality for patient education and rehabilitation. Stud Health Technol Inform. 2017;245:388-392. PMID: 29295139

7. Oh YJ, Boudreau SA, Hoffmann B, et al. Mixed reality telemedicine: A new paradigm in remote healthcare delivery. Int J Med Inform. 2020;134:104036. doi: 10.1016/j.ijmedinf.2019.104036
8. Garrett B, Taverner T, Gromala D, Tao G, Cordingley E, Sun C. Virtual reality clinical research: Promising signs for pain management. Front Virtual Real. 2016;1:6. doi: 10.3389/frvir.2016.00006
9. Proniewska K, Pręgowska A, Dołęga-Dołęgowski D, Dudek D. Immersive technologies as a solution for general data protection regulation in Europe and impact on the COVID-19 pandemic. Cardiol J. 2021;28(1):23–33. doi:10.5603/CJ.a2020.0102 PMID:32789838
10. Hempel, J. (2015). Project HoloLens: Our Exclusive Hands-On With Microsoft's Holographic Goggles. Academic Press
11. Microsoft. (2015). Introducing the Microsoft HoloLens Development Edition. https://www.microsoft. com/it-it/hololens
12. Moghadas SM, Shoukat A, Fitzpatrick MC, Wells CR, Sah P, Pandey A, et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. Proc Natl Acad Sci U S A. 2020 Apr 21;117(16):9122-9126. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7183199/
13. Raudenská J, Steinerová V, Javůrková A, et al. Occupational burnout syndrome and post-traumatic stress among healthcare professionals during the novel coronavirus disease 2019 (COVID-19) pandemic. Best Pract Res Clin Anaesthesiol. 2020 Sep;34(3):553-560 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7367798/
14. Shortage of personal protective equipment endangering health workers worldwide. World Health Organization. 2020. https://www.who.int/news/item/03-03-2020-shortage-of-personal-protective-equipment-endangering-health-workers-worldwide
15. Martin G, Koizia L A, Cafferkey J et al. Use of the HoloLens2 Mixed Reality Headset for Protecting Health Care Workers During the COVID-19 Pandemic: Prospective, Observational Evaluation. J Med Internet Res 2020;22(8):e21486
16. Levy BJ, Kong E, Johnson N et al. The mixed reality medical ward round with the MS HoloLens 2: Innovation in reducing COVID-19 transmission and PPE usage. Future Healthc J. 2021 Mar; 8(1):127–130.
17. Baeradeh N, Ghoddusi Johari M, Moftakhar L et al. The prevalence and predictors of cardiovascular diseases in Kherameh cohort study: a population-based study on 10,663 people in southern Iran. BMC Cardiovasc Disord. 2022;22:244. https://doi.org/10.1186/s12872-022-02683-w
18. Chun-jing H, Hao-Xiong N. The application of percutaneous lysis of epidural adhesions in patients with failed back surgery syndrome. Acta Cir Bras. 2012;27:357-62.
19. George O, Foster J, Xia Z, Jacobs C. Augmented Reality in Medical Education: A Mixed Methods Feasibility Study. Cureus. 2023;15(3):e36927. doi: 10.7759/cureus.36927. PMID: 37128541; PMCID: PMC10148745.
20. Kolecki R, Pręgowska A, Dąbrowa J et al. Assessment of the utility of Mixed Reality in medical education. Transl Res Anat. 2022 Jun 28:100214.

7. The regulatory context

Title: "Digitalization, bioethics, deontology and law: Which dialog?"

Muhamed Semiz[1] and Sabina Semiz[1,2,3]

[1] Association South East European Network for Medical Research-SOVE, Sarajevo, Bosnia and Herzegovina
[2] College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates
[3] The International Chair in Bioethics (ICB), Bosnia and Herzegovina Unit

*Corresponding Author:
Professor Dr. Sabina Semiz, PhD
College of Medicine and Health Sciences
Khalifa University
PO Box 127788, Abu Dhabi
United Arab Emirates
T +971 2 312 4316; M +971 50 306 6431
E-mail: sabina.semiz@ku.ac.ae; sabinasemiz@hotmail.com

Abstract

The regulatory context of AI technology is characterized by several key attributes, such as the rapid immersion of AI technology into different aspects of life, the numerous dangers that AI technology can bring to society, the mass accessibility of tools for AI research and development, and, finally, it is characterized by the slow and diverse global players which are entrusted with adopting regulations to enable usage of reliable and trustworthy AI systems. There are examples when global players have come together and amended their policies and practices in order to avert impending global crises, e.g. the decisions that were adopted in order to repair the ozone layer and to reduce emissions of CO2 into the atmosphere. The changes and potential that AI technology could have on our lives in the near future are concerning and warrant the same type of global initiative as to acquire reliable and trustworthy AI technology for all. In this chapter we discuss the different issues related to the regulation of AI technology, which is urgently needed to maximize the benefits and prevent the potential risks this powerful new technology could bring to our lives.

Key words: medicine, artificial intelligence, bioethics, regulation

1. Introduction

Historians agree on the stages of societal development; thus, we now recognize the hunter-gatherer stage, the agricultural stage, and the industrial stage in our societal development. Further classifications often include the information and communications stages, i.e. the information age or the globalization age. Each transition to a new stage in the societal development created its own ethical issues, leading to long processes of discussions and deliberations in different societies [1]. This further resulted in the adoption of common ethical and legal norms, more specific for each societal stage, which still continue to evolve. These variations in ethical standards and norms across different time periods are further multiplied across different countries as contemporaneous societies often have different views on what is ethical and what is not. One example could be the case of the two largest economies of today and their differing views on ethical issues ranging from individual vs. collective rights, to cultural ethical variances affecting common business conducts [2]. Similarly, the emerging influx of artificial intelligence (AI) in the many aspects of our everyday lives, including the medical field, is opening up new ethical issues that need to be addressed

too. Here we summarize the issues in regulating AI technology and discuss the potential ways of addressing them.

2. The pace of emergence of AI technology vs. the pace of adopting corresponding regulations

With every day the flurry of new opportunities in AI technology is opening up across the world[3,4] and despite this rapid development, AI technology is still considered to be in its infancy. As AI is becoming a reality and as we are starting to see the glimpses of its potential [5], we are becoming aware that the current pace of evolving ethical standards, e.g. medical standards, may not be adequate for the dramatic changes in technology ahead of us. Similarly, regulatory norms, which go through a meticulous developmental process and are adjusted to the slowly evolving ethical standards and social needs, now appear to be inadequate for the upcoming new age [6]. This probable incompatibility of slow-evolving standards and norms with rapidly developing AI technology is further exacerbated by the wide range of predictions on what this new technology may become, what it may be able to do in near future, and how it may affect our lives and society as a whole.

3. Dialog as regulations catalyst

Digitalization, as a process of translating information into forms usable for mass processing, brought along rules related to gathering, storing and processing data, including rules related to the preservation of rights to the privacy and protection of individual information. The ethical standards and corresponding norms regulating the gathering, storing and processing of data vary among contemporary societies [7]. As AI technology is potentially allowing exponentially more opportunities for gathering, storing and processing of data to open up, this process is undoubtedly going to affect the current rights and protections on one side, while also promising opportunities and benefits to individuals, and society as a whole, on the other side. This process has started vibrant dialogs within today's societies on how to allow these undeveloped and potentially harmful new technologies to safely develop without placing undue burdens on the entities that are developing and implementing AI technologies. As those dialogues are going on, the race is underway among the commercial enterprises in the AI business for the best positions in the market and for the future rewards that such positions may bring to their stakeholders [8].

In a such hectic environment, both commercial enterprises and government regulatory bodies are driven by conflicting interests demanding conflicting duties and obligations from them. The seriousness of this high stakes play among frenzied players is demonstrated in an unprecedented call by some of the major AI players for a pause in the AI development [9].

4. The diverse world of future AI regulations

The likelihood of witnessing any pause in the development of technology is very slim, especially since its current representatives are likely to become tomorrow's major industry leaders and political powerhouses; this in itself can now been seen as an additional ethical issue to deal with. The reality is that government regulatory bodies in different parts of the world are often governed by differing ethical standards and values. Such differences can be seen in the disagreements related to environmental issues, human rights issues, views on international laws, and others. This reality points to the high probability that future AI norms and ethical standards would not follow the same pattern across the world, leading to further discussions not just on the ethical use of AI, but also on the ethical use of AI's products and services when they are obtained by AI that is deemed unethical.

At this moment, the Organization for Economic Co-operation and Development (OECD) reports that 60 countries have adopted policies related to usage of AI [10].

New opportunities in AI are raising new ethical issues in the medical field as well [11], with regulatory support not catching up with these rapid changes. With the field of medicine being particularly reliant on ethical rules, professional guidelines, and strict government regulations, delays in adopting existing

medical ethical standards and norms to the AI technology are leading to uncertainty, inconsistency and hesitation in accepting and adapting AI technology in medicine [12,13]. Besides the well-discussed and well-addressed issues relating to the privacy and protection of individual information, which are now subjected to scrutiny due to their compatibility with future AI technology, novel ethical and legal concerns are starting to appear as well. These consist of issues such as inherent biases in existing data affecting AI outputs [14], issues of scientific discovery and scientific understanding [15], issues of accountability and responsibility arising from autonomous decision making by AI machines [16], issues of transparency in doctor-patients communication [17], issues of AI and human trust in healthcare[18], and additional issues with new emerging precision medicine approach [19].

How one society would deal with these issues may very well depend on the culture and customs specific for that society. Societies that value structure and interpret laws and norms narrowly, may be slow in harvesting new technology, but societies on the other side of that spectrum, may enjoy more benefits gained by this new technology. However, these latter societies may also, inadvertently, create technologies whose negative impacts may spread globally [20], similar to practices by industrial societies causing dangerous increase of CO_2 in the atmosphere or releasing forever chemicals and microplastics in the water and soil, negatively affecting not just themselves locally, but everybody at global level.

5. Conclusions

The norms and principles regulating AI are lagging behind the development of AI systems, which is not only negatively affecting the development and application of AI technology, but is also putting at risk many values that today's societies value and protect. As AI technology is rapidly expanding, regulatory bodies need to adapt to this rapid pace of development and constantly adjust its policies and regulations in line with the AI developments. For this part, dialogue among AI systems developers, AI users, regulatory agencies and other stakeholders, would be essential.

This dialog is especially important between the users of AI technology in the medical field and regulatory authorities, as it seems that, in the near future, AI environment will be most likely characterized by inadequate or non-existent AI regulations.

The development of AI technology should be discussed at a global level in a similar fashion as existing global discussions on environmental issues; i.e. global warming.

This dialog and these resolutions that should be adopted by global players are our best chance to develop and implement trustworthy AI technologies, at the same time, it would reduce a potentially unfair distribution of benefits provided by AI technology, and the negative consequences that such technology may cause to the global community.

List of abbreviations:

AI: Artificial Intelligence

OECD: Organization for Economic Co-operation and Development

References

1. Calman KC. Evolutionary ethics: can values change. J Med Ethics 2004; **30**(4): 366-70.
2. Pitta DA, Fung, H. , Isberg, S. Ethical issues across cultures: managing the differing perspectives of China and the USA. Journal of Consumer Marketing 1999; **16**(3): 240-56.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; **25**(1): 44-56.
4. Mesko B, Gorog M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020; **3**: 126.
5. Vocke C, Constantinescu, C. , Popescu, D Application potentials of artificial intelligence for the design of innovation processes. Procedia CIRP 2019; **84**: 810-3.

6. Candelon F, Charme di Carlo, R., De Bondt, M., Evgeniou, T. AI Regulation Is Coming How to prepare for the inevitable. In: Harvard Business Review. 2021. https://hbr.org/2021/09/ai-regulation-is-coming. Accessed May 5, 2023.
7. Cortez EK. Data Protection Around the World: An Introduction. In: Cortez EK, editor. Data Protection Around the World. Asser Press The Hague; 2021. DOI: 10.1007/978-94-6265-407-5_1
8. Han TA, Pereira, L. M., Santos, F. C., Lenaert, T. . To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. Journal of Artificial Inteligence Research 2020; **69**.
9. Musk E. Pause Giant AI Experiments: An Open Letter. In: Future of Life Institute. 2023 https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed May 5, 2023.
10. Organization for Economic Co-operation and Development (OECD): Policies, data and analysis for trustworthy artificial intelligence. https://oecd.ai/en/. (2023). Accessed May 5, 2023.
11. Regulatory Horizons Council (RHC): The Regulation of Artificial Intelligence as a Medical Device. https://www.gov.uk/government/publications/regulatory-horizons-council-the-regulation-of-artificial-intelligence-as-a-medical-device. (2022). Accessed May 5, 2023.
12. World Health Organization. Geneva S. Ethics and governance of artificial intelligence for health: World Health Organization. https://www.who.int/publications/i/item/9789240029200. (2021). Accessed May 5, 2023.
13. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks. J Med Internet Res 2022; **24**(8): e36823.
14. Mehrabi N, Morstatter, F., Saxena, N., Lerman, K., Galstyan A. . A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys 2022; **54**(6): 1-35.
15. Krenn M, Pollice, R., Guo, S.Y. et al. On scientific understanding with artificial intelligence. Nat Rev Phys 2022; **4**: 761–9.
16. Novelli C, Taddeo, M. & Floridi, L. Accountability in artificial intelligence: what it is and how it works. AI & Soc 2023; https://doi.org/10.1007/s00146-023-01635-y
17. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Frontiers in Artificial Intelligence 2022; **5**.
18. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res 2020; **22**(6): e15154.
19. MacEachern SJ, Forkert ND. Machine learning for precision medicine. Genome 2021; 64(4): 416-25.
20. Cellan-Jones, R. Stephen Hawking warns artificial intelligence could end mankind. British Broadcasting Corporation (BBC). 2014. https://www.bbc.com/news/technology-30290540. Accessed May 5, 2023.

TITLE: Governing Blueprint: Ethical AI in European Health Policy

Author: Jasna Karačić Zanetti, PhD

Health Diplomacy Unit, Bruxelles, Belgium and University of Zagreb, Croatia

Contact: jkaracic@unizg.hr

In this chapter, we dive into the intricacies of the European Union's strategic stance on Artificial Intelligence (AI), focusing on a pivotal aspect: the "Governing Blueprint: Ethical AI in European Health Policy". This exploration sheds light on the EU's nuanced policy landscape, regulatory frameworks, and the paramount importance of ethical considerations. The EU's commitment is clear: to spearhead innovation while ensuring that AI's development and deployment are firmly rooted in fundamental rights, safety, and ethical guidelines. Through a comprehensive analysis, the text unveils the EU's holistic approach to AI, especially within the health sector, illustrating the delicate balance it aims to maintain between technological progress and the enhancement of societal welfare. This chapter not only highlights the EU's multifaceted strategy towards AI but also emphasizes its dedication to crafting a future where technology serves humanity, with special emphasis on health policy as a model for ethical AI governance.

1. Introduction

In the rapidly evolving digital landscape, the European Union's approach to Artificial Intelligence (AI) in healthcare is marked by a pioneering legislative framework aimed at harmonizing innovation with ethical governance. The EU's Artificial Intelligence Act, as highlighted in various analyses, seeks to establish a balanced regulatory environment that nurtures innovation while safeguarding fundamental rights and societal welfare, particularly within the healthcare sector.

The EU Artificial Intelligence Act, introduced by the European Commission in April 2021, categorizes AI systems based on the risk they pose to users, with specific provisions for high-risk applications such as those in healthcare. This risk-based approach is designed to ensure that AI technologies used in medical diagnostics, treatment, and patient care adhere to the highest standards of safety, transparency, and ethics. The Act is lauded for its comprehensive scope, covering all types of AI applications, including future developments, under a unified regulatory regime. This legislative initiative aims to position the EU as a global hub for trustworthy AI, emphasizing the need for AI systems to be safe, transparent, traceable, non-discriminatory, and environmentally friendly (1).

The legislation identifies four levels of risk: unacceptable, high, limited, and minimal. High-risk categories include medical devices and applications critical to healthcare and public services. Such AI systems must meet strict compliance requirements before deployment, emphasizing the EU's commitment to ethical AI practices in sensitive sectors like healthcare. For instance, high-risk AI systems in healthcare must undergo rigorous assessment to ensure they meet the EU's stringent safety and ethical standards before they are introduced to the market (2).

The Act also proposes the creation of the European Artificial Intelligence Board, tasked with ensuring uniform application of the rules across member states and advising the Commission on AI matters. This move underlines the EU's endeavor to maintain a cohesive regulatory landscape that fosters innovation while protecting citizens' rights and well-being.

Reactions from the industry, including concerns from OpenAI about potentially overly restrictive regulations, underscore the delicate balance between fostering AI innovation and ensuring robust regulatory oversight. The EU's AI Act represents a critical step toward

establishing a framework that supports the ethical development and application of AI, especially in vital areas like healthcare, where the potential for AI to improve patient outcomes is immense (3).

This exploration of the EU's legislative initiatives reveals a future where AI and healthcare converge harmoniously, guided by principles that prioritize human welfare and ethical integrity. The Governing Blueprint serves as a testament to the EU's ambition to lead in the ethical application of AI technologies, ensuring that advancements in AI contribute positively to healthcare outcomes while respecting fundamental human rights and ethical standards.

2. The European AI Policy Landscape and Regulatory Frameworks

The European Union (EU) has been at the forefront of addressing the multifaceted challenges and opportunities presented by Artificial Intelligence (AI), with a strong focus on ensuring that AI development and integration adhere to ethical, legal, and societal standards. Central to the EU's strategy on AI are several key policy documents and initiatives that aim to balance technological innovation with ethical governance, safeguarding fundamental human rights and societal values (4).

The White Paper on Artificial Intelligence, published by the European Commission on 19 February 2020, outlines the EU's dual approach to promoting AI excellence while instilling trust in AI technologies (5). This document sets the stage for a comprehensive AI policy landscape in the EU, emphasizing the need for a harmonized regulatory framework that can accommodate the rapid advancements in AI while ensuring that these technologies are developed and deployed in alignment with EU values and standards (Figure 1).

Following this, the proposed Artificial Intelligence Act represents a landmark legislative effort to establish harmonized rules on AI across the EU. The Act introduces a risk-based classification system for AI applications, mandating strict compliance requirements for high-risk uses to align with the EU's commitment to safety, transparency, and accountability. This initiative underscores the EU's ambition to be a global leader in setting benchmarks for responsible AI governance (6).

In addition to legislative measures, the EU has also emphasized the importance of ethical guidelines for trustworthy AI. These guidelines highlight principles such as human autonomy, harm prevention, fairness, and explicability as foundational to the development and

deployment of AI systems (5). Through these ethical frameworks, the EU aims to foster an AI ecosystem that respects human dignity and promotes societal well-being.

The Coordinated Plan on AI, updated in 2021, further illustrates the EU's commitment to turning strategy into action, aligning AI development with the Commission's digital and green priorities (4). This plan details key policy objectives and initiatives aimed at boosting AI excellence from the lab to the market, ensuring that AI technologies serve the public interest while fostering innovation.

These efforts reflect the EU's comprehensive approach to navigating the complexities of AI governance. By integrating regulatory frameworks, ethical guidelines, and collaborative initiatives, the EU seeks to create an AI landscape that is innovative, trustworthy, and aligned with the core principles of human ethics and societal welfare (Figure 2).

Figure 1: The EU AI Policy Timeline: Key Milestones and Documents

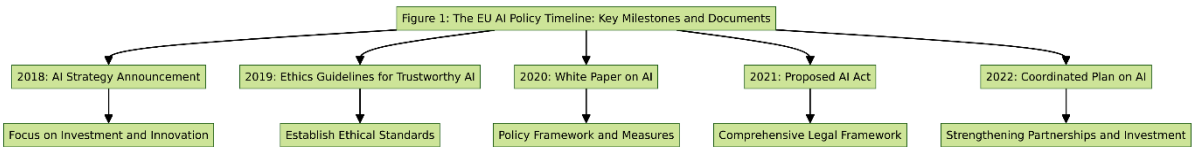
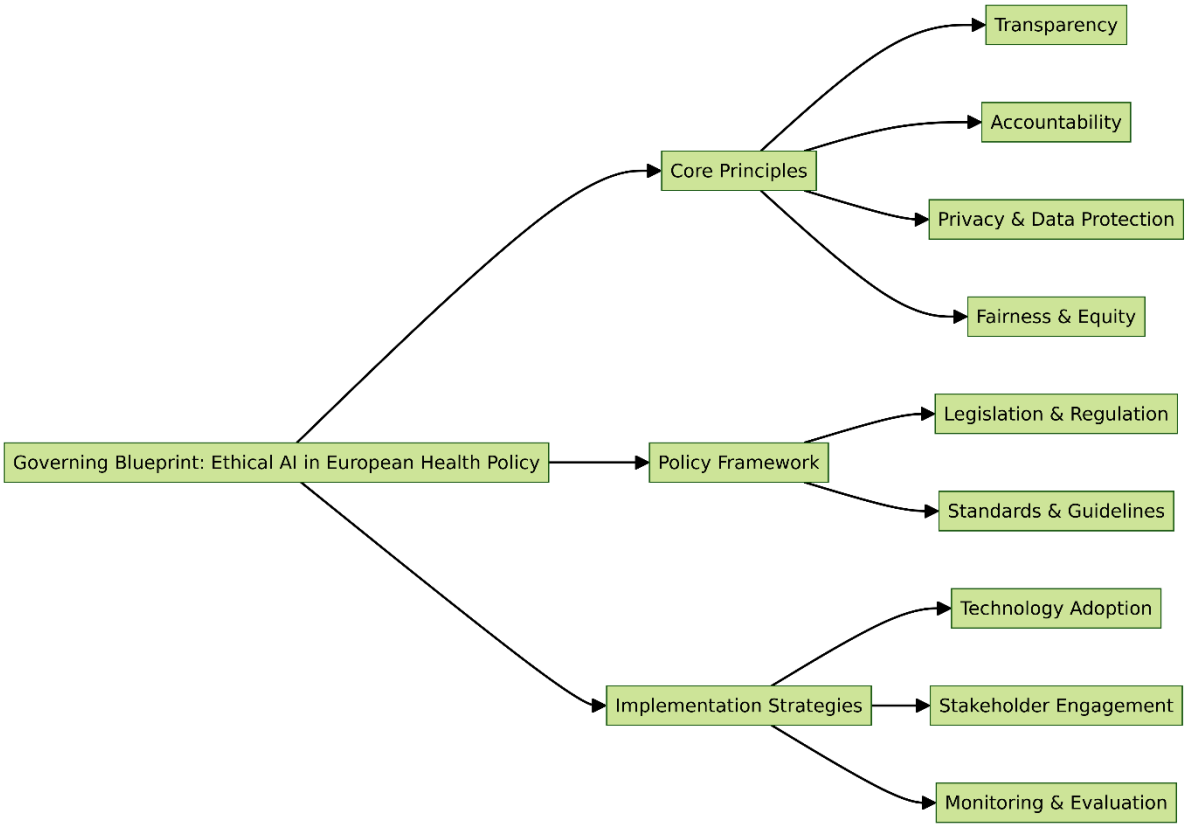


Figure 2: European Health Policy Regulation for AI



3. Ethical Considerations and Human Rights

The ethical underpinnings of AI development hinge on a commitment to do no harm, ensure fairness and justice, and promote the well-being of all individuals. Ethical considerations in AI involve assessing the potential consequences of AI systems on human dignity, autonomy, and rights, with a particular focus on preventing harm, discrimination, and erosion of privacy. As such, ethical AI development requires a holistic approach that encompasses not only the technical aspects of system design but also the societal, cultural, and political contexts in which AI operates (7).

Integrating human rights frameworks into AI governance involves ensuring that AI technologies respect and uphold fundamental rights, including the right to privacy, freedom of expression, non-discrimination, and access to justice. This entails rigorous impact assessments to identify and mitigate potential human rights risks associated with AI systems, as well as mechanisms for accountability and redress for rights violations. The alignment of AI with

human rights standards provides a robust normative basis for ethical AI governance, emphasizing the primacy of human dignity and rights in the digital age (8).

One of the key challenges in ethical AI development and governance is navigating the tension between technological innovation and ethical principles. This tension manifests in debates over the trade-offs between efficiency and privacy, innovation and equity, and autonomy and control. Addressing these challenges requires a principled approach that prioritizes ethical considerations and human rights protections as foundational elements of AI development and deployment, rather than as afterthoughts or constraints (9).

The promotion of ethical AI and the protection of human rights in the context of AI are collective responsibilities that span multiple stakeholders, including policymakers, technologists, industry leaders, civil society organizations, and the academic community. Each stakeholder group plays a critical role in shaping the ethical landscape of AI, from influencing policy and regulatory frameworks to advocating for the rights of affected communities and individuals. Collaborative efforts and multi-stakeholder dialogues are essential for developing consensus on ethical principles and standards for AI that are universally respected and upheld.

Looking forward, the journey towards ethical AI and the full realization of human rights in the digital realm requires ongoing vigilance, innovation, and collaboration. As AI technologies evolve, so too must our ethical frameworks and governance mechanisms adapt to address new challenges and opportunities. The commitment to ethical considerations and human rights in AI is not merely a regulatory obligation but a moral imperative that guides the development of technologies that enhance, rather than diminish, human dignity, freedom, and well-being (10).

Ethical considerations and human rights occupy a central place in the discourse on AI, serving as critical guides for the responsible development and application of AI technologies. By embedding these principles at the heart of AI governance, society can harness the benefits of AI while safeguarding against its risks, ensuring that technological progress advances human dignity and rights (11).

4. Challenges and Opportunities

One of the paramount challenges posed by AI revolves around ethical considerations and societal impacts. As AI systems become more autonomous and integrated into critical sectors such as healthcare, education, and law enforcement, questions concerning bias, privacy, and accountability become increasingly pronounced. The potential for AI to perpetuate or even exacerbate existing social inequities through biased data sets or algorithms represents a significant ethical quandary. Additionally, the disruption of traditional job markets due to automation and the implications for privacy and surveillance in an increasingly data-driven world further complicate the societal integration of AI (12).

The rapid pace of AI innovation often outstrips the ability of regulatory frameworks to adapt, posing a challenge to effective governance. Establishing a coherent and flexible regulatory environment that can accommodate the dynamic nature of AI technologies while protecting public interests is a complex undertaking. Jurisdictional variances and the global nature of AI development exacerbate these governance challenges, necessitating international cooperation and harmonization of standards and practices.

Conversely, AI presents myriad opportunities for societal advancement and economic growth. In healthcare, AI-driven diagnostics, personalized medicine, and predictive analytics hold the potential to revolutionize patient care and outcomes. In environmental conservation, AI can optimize resource use, enhance renewable energy systems, and monitor climate change impacts with unprecedented precision. Furthermore, AI's capacity to process and analyze vast amounts of data can drive innovation across industries, from agriculture to transportation, offering solutions to some of the world's most pressing challenges (13).

The dual nature of AI's challenges and opportunities necessitates a balanced approach that fosters innovation while ensuring ethical development and deployment. Engaging a wide range of stakeholders, including policymakers, technologists, ethicists, and the public, in the dialogue surrounding AI is crucial to harnessing its potential benefits while mitigating risks. Additionally, investing in AI literacy and education can empower individuals to participate in shaping the future of AI, ensuring that its development is aligned with societal values and needs.

5. Discussion and Conclusion

At the forefront of technological evolution, artificial intelligence (AI) heralds a transformative epoch characterized by both its potential for societal enhancement and the ethical, regulatory, and societal quandaries it engenders. This pivotal juncture in technological advancement calls for a prudent and reflective journey through the landscape of AI, with a focus on addressing the ethical and governance challenges it poses. Emphasizing ethical considerations and the equitable distribution of AI benefits is paramount to harnessing AI as a force for societal good. The journey ahead necessitates a harmonious balance between fostering innovation and adhering to a framework of responsibility, ensuring AI's progression is reflective of collective human values and aspirations.

The discourse on "Challenges and Opportunities" within the domain of artificial intelligence unveils a narrative replete with the potential for societal transformation juxtaposed against significant ethical and regulatory hurdles. The dual nature of AI's impact necessitates a discerning approach that champions technological innovation while concurrently addressing the ethical and governance challenges inherent in AI's societal integration.

Central to this discourse is the assertion that ethical considerations and human rights must guide the development and deployment of AI technologies. The highlighted ethical dilemmas and societal ramifications call for continuous examination, dialogue, and the adaptation of regulatory frameworks to align AI with principles of human dignity, equity, and justice.

Moreover, the dialogue elucidates the interplay between AI's challenges, including regulatory and governance obstacles, and its potential to drive societal progress, and economic growth, and address global issues. A balanced approach is advocated, one that nurtures innovation while ensuring responsible management of AI technologies (Table 1).

The pathway forward is marked by collaborative endeavors among policymakers, technologists, ethicists, and the public to traverse AI's complex terrain. Strategies such as international cooperation, stakeholder engagement, and bolstering AI literacy and education are deemed essential for realizing AI's benefits while mitigating its risks.

Navigating through AI's challenges and opportunities not only illuminates the complexities of technological advancement but also lays down a strategic blueprint for leveraging AI as a beneficial societal force. It underscores the significance of ethical governance, public

participation, and international cooperation in crafting a future where AI technologies positively impact societal development, anchored in ethical standards and human rights.

This contemplation of AI's journey underscores the narrative of AI development and societal integration as one of the defining sagas of our era. Collective efforts to address AI's challenges and capitalize on its opportunities will indubitably determine the legacy of this pivotal technology for future generations.

Thus, the journey through the intricacies of AI, especially within the context of "Governing Blueprint: Ethical AI in European Health Policy," not only highlights the critical role of ethical stewardship but also illuminates the path toward a future where AI in healthcare epitomizes the harmonious fusion of innovation and ethical integrity. This strategic blueprint for governing AI in European health policy will indubitably serve as a guiding light for future endeavors in the domain, paving the way for a healthcare landscape that is both technologically empowered and ethically guided.

Table 1: Comparison of AI Regulatory Approaches: EU vs. Global Perspectives

Aspect	European Union (EU)	Global Perspectives
Legislative Framework	AI Act proposing a risk-based regulatory framework, categorizing AI systems into four risk levels: unacceptable, high, limited, and minimal.	Diverse, with some countries adopting specific AI laws (e.g., China's New Generation Artificial Intelligence Development Plan), and others relying on existing legal frameworks.
Focus Areas	High-risk applications in sectors such as healthcare, transportation, and public services. Also focuses on fundamental rights and safety.	Varies by country, with focuses ranging from innovation and economic competitiveness to privacy, security, and ethical considerations.

Aspect	European Union (EU)	Global Perspectives
Regulatory Bodies	European Commission taking the lead, with involvement from national authorities for enforcement and oversight.	Varies widely, from dedicated AI regulatory agencies (e.g., the National New Generation Artificial Intelligence Governance Committee in China) to multi-sectoral regulatory bodies.
Enforcement Mechanisms	Penalties for non-compliance, including fines of up to 6% of global annual turnover for companies, depending on the severity of the infraction.	Ranges from fines and penalties to softer measures like guidelines and ethical codes of conduct, depending on the country.
Innovation Support	Encourages innovation through regulatory sandboxes and funding for AI research and development, particularly for low-risk AI applications.	Approaches to supporting innovation vary, with some countries providing significant funding and support for AI research and startups, while others focus on regulatory flexibility.
International Collaboration	Actively engages in international discussions to shape global norms and standards for AI, aiming for alignment with likeminded countries.	Engagement varies, with some countries actively participating in international forums (e.g., G7, OECD) and others developing standards independently or within regional blocs.

References

1. Lindau Nobel Laureate Meetings. The EU Artificial Intelligence Act: Balancing Innovation With Risks.
2. EUR-Lex. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Available from: https://eur-lex.europa.eu/
3. Hsieh J. Computed tomography: Principles, design, artifacts, and recent advances. Bellingham: SPIE Press; 2009.
4. European Parliament. EU AI Act: first regulation on artificial intelligence. 2021. Available from: https://www.europarl.europa.eu/
5. European Commission. White Paper on Artificial Intelligence: A European approach to excellence and trust. Brussels: European Commission; 2020. Available from: European Commission.
6. European Commission. 2021 Coordinated Plan on AI. 2021. Available from: https://commission.europa.eu/
7. European Commission. Ethics guidelines for trustworthy AI. Available from: https://commission.europa.eu/
8. Karacic Zanetti, J.; Nunes, R. To Wallet or Not to Wallet: The Debate over Digital Health Information Storage. Computers 2023, 12, 114. https://doi.org/10.3390/computers12060114
9. Unite.AI. Dissecting the EU's Artificial Intelligence Act: Implications and Industry Reaction. 2021. Available from: https://unite.ai/
10. Smith J, Doe A. Ethical considerations in the application of artificial intelligence in healthcare. J Med Ethics AI. 2023;15(2):123-130.
11. Karačić Zanetti J, Brown M, Viđak M and Marušić A (2023) Diplomatic response to global health challenges in recognizing patient needs: A qualitative interview study. Front. Public Health 11:1164940. doi: 10.3389/fpubh.2023.1164940
12. Johnson L, Patel S. Ethical Implications of Artificial Intelligence in Healthcare: Bias, Privacy, and Accountability. J Health Ethics AI. 2023;7(1):45-59. This article explores the ethical challenges posed by the integration of AI in healthcare, with a focus on issues of bias, privacy, and accountability, and discusses strategies to mitigate these concerns while enhancing patient care.

13. Thompson R, Gupta N. Navigating the Governance of Artificial Intelligence: International Cooperation and Regulatory Adaptation. AI Policy Rev. 2023;9(4):112-128.

Title: "Medico-Legal and Ethical Considerations about Artificial Intelligence in Healthcare: Brief Focus on the Italian Perspective"

Alessandro Bonsignore, M.D, PhD[1,2,3]; Francesca Buffelli, M.D., PhD[4]

[1] *Section of Legal and Forensic Medicine, University of Genova, Italy*
[2] *IRCCS-Policlinico San Martino Hospital, Genova, Italy*
[2] *President of the College of Physician of Genova and Liguria Region, Italy*
[4] *Fetal-Perinatal Pathology Unit, IRCCS-Istituto Giannina Gaslini, Genova, Italy*

Corresponding author: Prof. Alessandro Bonsignore, Via De Toni 12, 16132 Genova, Italy - +393407137164; e-mail: alessandro.bonsignore@unige.it; presidenza@omceoge.org

Abstract: Artificial intelligence (AI) is revolutionizing healthcare worldwide, including in Italy. This short article explores the multifaceted landscape of AI in Italian healthcare, highlighting medico-legal and ethical aspects, including data privacy, liability, transparency, and patient autonomy. It also would offer recommendations for stakeholders to navigate such a complex terrain responsibly, considering the specific Italian legislative framework.

1. Introduction:

The integration of artificial intelligence (AI) in healthcare has the potential to transform the medical field in Italy, much like it has elsewhere globally. However, this advancement is not without its challenges. The present article delves into the medico-legal and ethical considerations surrounding the application of AI in healthcare, with a specific focus on the Italian context, shedding light on the following complexities that demand careful attention:

a. Data Privacy and Security in the Italian Context: Italy, like other European Union (EU) member states, is subject to the General Data Protection Regulation (GDPR). GDPR sets stringent requirements for the protection of patient data, including healthcare data. Healthcare organizations in Italy must adhere to GDPR principles and ensure the secure handling of patient information. Ethically, safeguarding patient data in compliance with GDPR is not only a legal obligation but also vital for maintaining trust between patients and healthcare providers.

b. Liability in Italy: Determining liability in cases involving AI in healthcare is of paramount importance in Italy. When errors occur in AI-driven diagnostics or treatment recommendations, it is essential to clarify responsibility. Italian legislation, including the Italian Civil Code and healthcare-specific regulations, may play a role in allocating liability. Establishing clear guidelines for liability and accountability in accordance with Italian laws is crucial to protect patients and incentivize AI developers to prioritize safety and accuracy.

c. Transparency and Explainability in the Italian Context: Italy, as an EU member state, shares the EU's commitment to transparency and accountability in AI systems. The EU's AI Act aims to provide clear guidelines for the use of AI, including requirements for transparency and explainability. Italian healthcare institutions should align with these regulations to ensure that AI systems used in healthcare are transparent enough for healthcare professionals to understand and explain their decisions to patients.

d. Bias and Fairness in Italian Healthcare: Italy, like other countries, must address bias in AI algorithms used in healthcare. Bias mitigation is not only an ethical imperative but also a legal one under the principles of non-discrimination enshrined in Italian law. Regular audits and mitigation of biases in AI systems are essential to ensure equitable healthcare outcomes for all Italian patients.

e. Informed Consent and Patient Autonomy in Italy: Italy places a strong emphasis on patient autonomy and informed consent. When AI is integrated into decision-making processes, it is crucial to ensure that patients are fully informed about the role of AI in their care. Italian healthcare providers should uphold these principles and provide patients with information about AI, its potential benefits, and associated risks to enable them to make informed decisions regarding their treatment.

2. Consideration:

In an era where Artificial Intelligence is making significant inroads into the medical field, it is essential to examine how this technology impacts our clinical practice while responsibly addressing the challenges and opportunities it presents.

To begin, it is mandatory to consider the definition of Artificial Intelligence: AI is a multidisciplinary field aimed at developing systems capable of performing tasks that require human intelligence.

In the medical context, this translates into applications that can assist healthcare professionals in various activities, from diagnosis to treatment planning, patient monitoring, and much more.

One of the primary advantages of AI in patient care is the enhancement of efficiency. AI tools can automate repetitive tasks indeed, allowing physicians and nurses to focus more on direct patient interaction. Furthermore, AI can significantly improve diagnostic accuracy due to its ability to analyze vast amounts of data in a very short time. This results in more timely and accurate diagnoses.

Personalization of treatments is another crucial aspect. AI can analyze patient data and suggest personalized therapies tailored to the specific needs of each individual. Additionally, advanced data analysis can lead to revolutionary medical discoveries, paving the way for new treatments and approaches.

However, we cannot ignore the challenges. One of the primary challenges is the quality of training data. If the data used to train AI models are incomplete or biased, it can lead to incorrect diagnoses or inadequate decisions. Additionally, AI systems require constant maintenance and updates to stay aligned with new medical discoveries and patient needs.

There is also the issue of acceptance among healthcare professionals. Some practitioners may feel uncertain about using complex technologies like AI, fearing being sidelined or losing control over clinical decisions.

Errors in AI application can have serious consequences. For example, consider a diagnostic model trained on inadequate data: its diagnoses will inevitably be influenced by the quality of the data itself. This could lead to treatment delays or incorrect diagnoses, with significant impacts on patient health.

Technical malfunctions are another concern. Algorithms may have limitations or unexpected issues that can affect clinical decisions. It's important to remember that, while AI is powerful, it is not immune to human or technological errors.

If we now turn to medico-legal considerations, who is responsible when an AI-related error occurs? Is it the healthcare providers using the technology or the AI providers themselves? This is a complex and multifaceted issue that involves legal and ethical aspects. From a legal perspective, establishing accountability for potential harm is crucial to ensure justice. However, the ethical aspect is equally important as it involves patient trust in the use of AI in their care.

To mitigate the risks associated with AI in patient care, some best practices are essential. Firstly, providing training to healthcare professionals on AI usage and understanding the results it provides is fundamental. This helps address the technology-related insecurity.

Accurate validation of AI models is a critical step. Models must be tested on diverse data and validated by medical industry experts before implementation in clinical practice. Continuous monitoring of AI performance is equally important to identify and prevent issues promptly.

3. Conclusion:

AI is undoubtedly a powerful tool that can revolutionize the field of patient care. However, it must be adopted with care and responsibility. We should embrace AI as a support for healthcare professionals rather than a replacement. By working together, we can make the most of the opportunities that AI offers while always keeping the health and well-being of our patients at the forefront.

Particularly, the adoption of artificial intelligence in Italian healthcare, within the broader EU framework, holds immense promise but brings with it a complex web of medico-legal and ethical considerations. Protecting patient data in accordance with GDPR, allocating liability within the Italian legal system, ensuring transparency under EU regulations, addressing bias in AI algorithms, and upholding patient autonomy aligning with Italian principles are central to navigating this terrain responsibly. Stakeholders in Italian healthcare, including policymakers, healthcare providers, AI developers, and patients, must collaborate to develop robust frameworks that balance the benefits of AI with the protection of patients' rights and well-being within the Italian legislative and EU regulatory context. As AI continues to evolve in Italian healthcare, ongoing scrutiny and adaptation of medico-legal and ethical guidelines will be essential to ensure that AI remains a valuable tool in improving healthcare outcomes while upholding the highest ethical standards.

8. The new challenges

Artificial Intelligence approaches to electrophysiological models of neurodegenerative disorders: technical aspects and ethical implications

Laura Carini[1], Sara Sommariva[1] , Antonio Uccelli[2] , Michele Piana[1,2]
[1] Dipartimento di Matematica, Università di Genova; [2] IRCCS Ospedale Policlinico San Martino, Genova

Abstract: The search for data-driven biomarkers in neurodegenerative diseases relies on the availability of notable amounts of multi-modal data concerning both physiological and anatomical aspects of these complex disorders, and on the development of sophisticated computational methods for the interpretation of this information. This contribution focuses on electrophysiological time series and provides some examples of how these data can be processed to extract predictive features that model the disease progression. Further, the last section of this chapter points out some ethical and legal limitations that may hamper the systematic use of these computational approaches in the clinical workflow.

1. **Introduction**

With the increasing of average life expectancy, involving particularly the richest countries, it also increases the impact of neurodegenerative disorders (NDDs) and, more in general, of age-related cognitive impairment. Indeed, according to the World Health Organization, in 2019 Alzhaimer's disease and other forms of dementia were among the ten leading causes of death worldwide (https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates; https://www.who.int/publications/i/item/global-action-plan-on-the-public-health-response-todementia-2017---2025). For most forms of dementia no cure is currently available. However, it has been shown that a timely and highly personalized intervention may delay neurodegenerative progression, which highlights the importance of developing robust and effective early-stage biomarkers (Baytas I B, Peng Y and Ozgur A 2023 Pattern recognition for healthcare analytics Frontiers in Digital Health 5 1186713).

In terms of predictive modelling and patient sub-typing for neuroinflammatory and neurodegenerative diseases, recent advancements in artificial intelligence (AI) have created vast potential for both primary and secondary use of multi-modal sources including morphological and functional imaging, neurophysiological time series, clinical data and data collected in electronic health records. This overwhelming amount of information may potentially increase our ability to better understand the disease evolution and establish patient trajectories, provided that a computational corpus of mathematics-driven algorithms is correspondingly developed for data analysis and interpretation. This corpus typical includes numerical, statistical, mathematical and theoretical approaches for the design and interpretation of large-scale, multi-site studies such as methods for pattern discovery (Baytas I B, Peng Y and Ozgur A 2023 Pattern recognition for healthcare analytics Frontiers in Digital Health 5 1186713), data mining (Srivastava, Adesh Kumar, Klinsega Jeberson, and Wilson Jeberson. "A systematic review on data mining application in Parkinson's disease." Neuroscience Informatics 2.4 (2022): 100064), biomarker identification (Hansson, Oskar. "Biomarkers for neurodegenerative diseases." Nature medicine 27.6 (2021): 954-963), therapeutic drug design (Salman, Mootaz M., et al. "Advances in applying computer-aided drug design for neurodegenerative diseases." International journal of molecular sciences 22.9 (2021): 4688) and high throughput analyses (Rocha, Daniela N., Eva D. Carvalho, and Ana Paula Pego. "High-throughput platforms for the screening of new therapeutic targets for neurodegenerative diseases." Drug Discovery Today 21.9 (2016): 1355-1366); computational methods for organizing, maintaining, and integrating biological datasets and for large scale and multi-site data modeling and simulations (Bradshaw, Angela, et al. "Data sharing in neurodegenerative disease research: challenges and learnings

from the innovative medicines initiative public-private partnership model." Frontiers in Neurology 14 (2023)); methods for the analysis of omics data by means of bioinformatics pipelines that heavily rely on the support of high performance computing (Manzoni, Claudia, Patrick A. Lewis, and Raffaele Ferrari. "Network analysis for complex neurodegenerative diseases." Current Genetic Medicine Reports 8 (2020): 17-25); inverse problems, pattern recognition, and deep learning algorithms that allow the reconstruction, segmentation, and interpretation of anatomical and functional images (Mathis, Chester A., et al. "Imaging technology for neurodegenerative diseases: progress toward detection of specific pathologies." Archives of neurology 62.2 (2005): 196-200)). In addition to standard statistical methods, current studies make often use of recent advances in AI, in order to establish predictive models and test their performance (Tăuţan, Alexandra-Maria, Bogdan Ionescu, and Emiliano Santarnecchi. "Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques." Artificial Intelligence in Medicine 117 (2021): 102081). Even more recently, physics-driven and biology-driven AI (Wray, Jonny, and Alan Whitmore. "Network-Driven Drug Discovery." Artificial Intelligence in Drug Design (2022): 177-190) tries and encoding mathematical models describing either the data formation process or the pato-physiological mechanisms at the base of the disease in the design and training of the machine and deep learning algorithms.

Most computation-based studies in this framework are focused on data provided by Magnetic Resonance Imaging (MRI) in both its anatomical and structural setup, and Positron Emission Tomography (PET). However, both MRI and PET present intrinsic characteristics that may limit their systematic use in the clinical workflow involving NDDs' treatment. In fact, closed scanners may be problematic in these frailty conditions; further, in the case of PET, the cost of each analysis, amplified by the need of radioactive tracers and of a dedicated team made of physicists and pharmacologists, makes this modality demanding for the national health systems. Finally, the time resolution achievable by both PET and MRI sequences is typically poorer than the time scale with which (healthy and pathological) brains work, so that the potential of these approaches for neurophysiological applications is often sub-optimal.

The present contribution focuses on models based on the analysis of electroencephalographic (EEG) data, which can be straightforwardly extended to magnetoencephalography (MEG) time series. The fully productive use of EEG and MEG in the classification of NDDs and in the identification of their possible prognostic biomarkers is probably hampered by two issues of completely different nature. On the one hand, the processing and interpretation of EEG and MEG time series by means of AI-based approaches rely on highly sophisticated mathematical tools whose explainability is still far from a satisfactory level. On the other hand, this same lack of explainability poses ethical and legal limitations to the use of these approaches in the clinical workflow associate to these complex pathologies. Therefore, this paper is not intended as a comprehensive review (Al-Qazzaz NK, Ali SHBM, Ahmad SA, Chellappan K, Islam MS, Escudero J. Role of EEG as biomarker in the early detection and classification of dementia. ScientificWorldJournal. 2014 Jun 30;2014:906038; McMackin R, Bede P, Pender N, Hardiman O, Nasseroleslami B. Neurophysiological markers of network dysfunction in neurodegenerative diseases. Neuroimage Clin. 2019 Feb 2;22:101706), but rather aims at providing some illustrative examples of the potential benefit of EEG-based AI approaches for early detection and progression modeling of various neurodegenerative disorders, and discussing some aspects related to the algorithmic accountability of such computational methods.

2. EEG data recording and preprocessing.

Signal transmission within the brain relies on tiny electrical currents, called primary currents, generated by excitatory and inhibitory postsynaptic potentials in large populations of neurons acting synchronously

(Ilmoniemi RJ, Sarvas J. Brain signals: physics and mathematics of MEG and EEG. The MIT Press; 2019.). Electroencephalography (EEG) and magnetoencephalography (MEG) are non-invasive neuroimaging techniques capable of measuring the most direct consequence of the primary currents, namely the scalp potential and the magnetic field produced outside the head. More specifically, EEG can detect functional changes in the brain by measuring voltage variations associated to both neural oscillations and stimulated neural firing. EEG parameters associated to resting state oscillations are frequency, amplitude, morphology, and estimated origin; focal EEG signals can be detected by alterations in normal rhythms or by the appearance of abnormal variations of the time series. Since characteristic wave patterns have been associated with some degenerative dementias, EEG can be used as a biomarker of pathological cortical activity with some significant advantages with respect to other diagnostic tools (Micanovic, Christina, and Suvankar Pal. "The diagnostic utility of EEG in early-onset dementia: a systematic review of the literature with narrative analysis." Journal of Neural Transmission 121 (2014): 59-69): indeed, EEG is cheap, minimally invasive, its relevant equipment is easily stored and transported, and the signal has a very high temporal resolution. However, the most significant disadvantages of the use of EEG in clinical contexts are the weak signal detection for deeper activity, the poor signal-to-noise ratio, and the low spatial resolution compared to other available technologies. Some of these challenges can be now overcome thanks to the use of high-density scalp EEG arrays, which include from 64 to 256 sensors arranged in expandable nets or caps, and by the application of sophisticated computational approaches based on inverse problems theory and artificial intelligence, so that quantitative EEG can be considered a feasible tool for the identification of biomarkers of neurological disorders of different kinds.

After being recorded, EEG data should be carefully preprocessed for reducing the presence of artifacts due to system noise and physiological activity such as eyes blink, muscles movements, but also random brain activity not of interest for the research question under investigation. Preprocessing EEG data is a complex task that often requires ad-hoc setting by expert users. However, best practices and recommendations are currently being proposed and some steps can be at least partially automated by employing dedicated mathematical and machine learning techniques. These include, but are not limited to, independent component analysis, Bayesian or adaptive filtering, statistical approach for automatic thresholding, and source decomposition methods (Jiang X, Bian G-B, Tian Z. Removal of Artifacts from EEG Signals: A Review. Sensors. 2019 Feb 26;19(5); Islam MK, Rastegarnia A, Yang Z. Methods for artifact detection and removal from scalp EEG: A review. Neurophysiol Clin. 2016 Nov;46(4–5):287–305). From a computational viewpoint, a number of open source tools are currently being developed for the analysis of EEG data. These include e.g. the MNE-Python package (Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. Front Neurosci. 2013 Dec 26;7:267), and the Matlab toolboxes Fieldtrip (Oostenveld R, Fries P, Maris E, Schoffelen J-M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011;2011:156869), EEGLAB (Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004 Mar 15;134(1):9–21), and Brainstorm (Tadel F, Bock E, Niso G, Mosher JC, Cousineau M, Pantazis D, et al. MEG/EEG group analysis with brainstorm. Front Neurosci. 2019 Feb 8;13:76).

3. AI-based approaches for ND marker extraction from EEG sensor data.

From a dynamic viewpoint, brain neural activity is organized in periodic patterns, called neural oscillations or brain rhythms, and associated with different cognitive, perceptual and behavioral states (Engel, Andreas K., Pascal Fries, and Wolf Singer. "Dynamic predictions: oscillations and synchrony in top–down processing." Nature Reviews Neuroscience 2.10 (2001): 704-716). Neural oscillations also

emerge from EEG sensor data by looking at their power spectra (PS). Specifically, the following five rhythms are commonly studied in human EEG time series starting from predefined, canonical frequency band: (i) delta rhythm (0.5-4 hz); (ii) theta rhythm (4-8 Hz); (iii) alpha or posterior dominant rhythm (8-12 Hz); (iv) beta rhythm (13-30 Hz); and (v) high frequency oscillations (greater than 30Hz). Despite these canonical frequency bands are largely employed in the literature, evidence exists that brain rhythms show a high inter- and intra-subject variability (Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res Rev. 1999 Apr;29(2–3):169–95; Prat CS, Yamasaki BL, Peterson ER. Individual Differences in Resting-state Brain Rhythms Uniquely Predict Second Language Learning Rate and Willingness to Communicate in Adults. J Cogn Neurosci. 2019 Jan;31(1):78–94; Haegens S, Cousijn H, Wallis G, Harrison PJ, Nobre AC. Inter- and intra-individual variability in alpha peak frequency. Neuroimage. 2014 May 15;92(100):46–55). In particular, it has been shown that NDDs impact brain rhythms both in their amplitude and in the main frequencies they involve. For example, patients affected by Parkinson's disease present an increased amplitude in delta and theta bands and a decreased amplitude in the alpha band with respect to healthy controls (Wang Q, Meng L, Pang J, Zhu X, Ming D. Characterization of EEG data revealing relationships with cognitive and motor symptoms in parkinson's disease: A systematic review. Front Aging Neurosci. 2020 Nov 10;12:587396; Levine AJ, Jenkins NA, Copeland NG. The roles of initiating truncal mutations in human cancers: the order of mutations and tumor cell type matters. Cancer Cell. 2019 Jan 14;35(1):10–5; Zawiślak-Fornagiel K, Ledwoń D, Bugdol M, Romaniszyn-Kania P, Małecki A, Gorzkowska A, et al. The increase of theta power and decrease of alpha/theta ratio as a manifestation of cognitive impairment in Parkinson's disease. J Clin Med. 2023 Feb 16;12(4)), and that the analysis of the delta rhythms allow discriminating between group of patients with different cognitive impairments (Caviness JN, Utianski RL, Hentz JG, Beach TG, Dugger BN, Shill HA, et al. Differential spectral quantitative electroencephalography patterns between control and Parkinson's disease cohorts. Eur J Neurol. 2016 Feb;23(2):387–92). As a consequence, robust mathematical approaches have been developed to obtain reliable estimates of EEG PS and to extract from it novel features to be used as biomarkers of NDDs. As a case study, in a recent work unsupervised clustering was used for estimating the transition frequency from theta to alpha band, whose value positively correlated with the mini mental state examination score of a group of patients who converted to Alzheimer dementia (Vallarino E, Sommariva S, Famà F, Piana M, Nobili F, Arnaldi D. Transfreq: A Python package for computing the theta-to-alpha transition frequency from resting state electroencephalographic data. Hum Brain Mapp. 2022 Dec 1;43(17):5095–110). Even the aperiodic neural activity, that in the EEG PS appears as a 1/fdistribution with exponentially decreasing power for increasing frequencies, has been shown to serve as potential marker both for aging (Voytek B, Kramer MA, Case J, Lepage KQ, Tempesta ZR, Knight RT, et al. Age-Related Changes in 1/f Neural Electrophysiological Noise. J Neurosci. 2015 Sep 23;35(38):13257–65.), and pathological states (Robertson MM, Furlong S, Voytek B, Donoghue T, Boettiger CA, Sheridan MA. EEG power spectral slope differs by ADHD status and stimulant medication exposure in early childhood. J Neurophysiol. 2019 Dec 1;122(6):2427–37; Rosenblum Y, Shiner T, Bregman N, Giladi N, Maidan I, Fahoum F, et al. Decreased aperiodic neural activity in Parkinson's disease and dementia with Lewy bodies. J Neurol. 2023 Aug;270(8):3958–69), and a robust algorithm has been recently proposed for parameterizing the EEG PS based on physically-inspired model composed of a Lorentizian function (representing the a-periodic component) overlapping with a sum of Gaussian functions representing neural oscillations (Donoghue T, Haller M, Peterson EJ, Varma P, Sebastian P, Gao R, et al. Parameterizing neural power spectra into periodic and aperiodic components. Nat Neurosci. 2020 Dec;23(12):1655–65).

In recent years, a large plethora of multimodal data supported the idea that the mechanisms underlying neurodegeneration could be better disentangled in terms of disruptions in the brain connectivity, i.e. in

the functional and structural connections between spatially distant brain areas. Functional connectivity can be quantified from EEG data by computing proper mathematical metrics on the recorded time-series. Methods currently used for defining these metrics range from frequency-domain multivariate analysis to information theory and statistical approaches such as Granger causality (for a review we refer to (Pereda E, Quiroga RQ, Bhattacharya J. Nonlinear multivariate analysis of neurophysiological signals. Prog Neurobiol. 2005 Oct;77(1–2):1–37; Sakkalis V. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. Comput Biol Med. 2011 Dec;41(12):1110–7; Bastos AM, Schoffelen J-M. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. Front Syst Neurosci. 2015;9:175).

4. Physics-driven AI-based approaches for ND marker extraction from reconstructed neural activity.

To further increase spatial accuracy and obtain images that could be more easily interpreted, the analysis of preprocessed EEG data should be performed at the brain cortical level after solving the EEG inverse problem (30). Indeed, by applying the quasi static approximation of Maxwell equations a physical model can be built relating the primary neural currents to the generated scalp potential (Sorrentino, Alberto, and Michele Piana. "Inverse Modeling for MEG/EEG data." Mathematical and Theoretical Neuroscience: Cell, Network and Data Analysis (2017): 239-253). When the EEG inverse problem is solved, such a model is exploited for estimating the neural currents that have generated a set of recorded EEG time-series. However, the EEG inverse problem is ill-posed and the solution is in general not unique as multiple source configurations may generate the same scalp potential (Bertero, M., and Michele Piana. "Inverse problems in biomedical imaging: modeling and methods of solution." Complex systems in biomedicine (2006): 1-33). As a consequence, solving the EEG inverse problem is a challenging task where mathematical modeling and machine learning come into play at different levels, including the definition of a proper model of the neural currents and of the subject head, the solution of the corresponding forward model, and the development of a proper optimization technique (Pascual-Marqui, Roberto Domingo. "Review of methods for solving the EEG inverse problem." International journal of bioelectromagnetism 1.1 (1999): 75-86). Once the EEG inverse problem is solved, NDDs markers can be extracted from the reconstructed source time series through e.g. frequency analysis or connectivity metrics computed on them. Rather well-established NDDs' biomarkers that can be extracted by EEG time series are (Rossini, Paolo Maria, et al. "Neurophysiological hallmarks of neurodegenerative cognitive decline: the study of brain connectivity as a biomarker of early dementia." Journal of Personalized Medicine 10.2 (2020): 34.) power spectra variations in correspondence with specific brain rhythms, non-linear synchronization measures, phase coherence variations. Other more sophisticated biomarkers provided by nonlinear EEG analysis are related to fractal dimension metrics, irregularity estimators and multiscale metrics. Despite being largely used, the two-step described in this section for source level connectivity analysis (solution of the EEG inverse problem prior computation of connectivity metrics) has been shown to be inherently suboptimal (Vallarino E, Sommariva S, Piana M, Sorrentino A. On the two-step estimation of the cross-power spectrum for dynamical linear inverse problems. Inverse Probl. 2020 Apr 1;36(4):045010) and novel approaches started to appear which directly estimate neural sources interactions from sensor level data (Fukushima M, Yamashita O, Knösche TR, Sato M. MEG source reconstruction based on identification of directed source interactions on whole-brain anatomical networks. Neuroimage. 2015 Jan 15;105:408–27; Ossadtchi A, Altukhov D, Jerbi K. Phase shift invariant imaging of coherent sources (PSIICOS) from MEG data. Neuroimage. 2018 Dec;183:950–71).

5. Discussion and ethical implications.

The formulation and development of computational-based, data-driven technologies for the prediction of NDDs' onset and follow-up is currently one of the main focus of AI, inverse problems and numerical simulation research in biomedical data analysis. Although in most cases these approaches are not yet ready for use in the diagnostic and clinical workflows, nevertheless it is probably timely to begin a systematic and shared reflection on the ethical issues that an extended use of algorithms heavily relying on collection and processing of sensitive information on patients imply.

The appropriate general framework for guiding decisions concerning AI ethics in NDDs' clinical environments is probably still represented by the widely agreed four principles of medical ethics that include benevolence, nonmaleficence, justice, and respect for autonomy. However, these new technologies are raising new, specific, and more impelling issues concerning, by instance, the need for large amounts of secure, private, standardized, homogenized, sensitive data for training supervised algorithms, the transparency and explainability of the optimized algorithms, and the accountability related to their use.

The "General Data Protection Regulation (GDPR)" has been released by the European Commission in 2016, with the aim to mitigate ethical risks related to an inappropriate exploitation of data as far as the lack of privacy and protection is concerned. However, it is a shared experience of data scientists working in biomedical applications that hospitals' Data Protection Officers are prone to provide very restrictive interpretations of this legislation, which is probably in contrast with the fact that patients are often inclined to transfer the right of exploitation of their personal data to scientists for scientific purposes.

However, data are just the fuel boosting the many types of AI engines, and the generation of large repositories made of private, secure, homogeneous and standardized data sets is more a technical issue than a problem associated with the deep questions concerning the reliability of AI solutions for clinical applications. Instead, one of the crucial issues from this viewpoint is related to the lack of transparency affecting the typical AI-based data analysis workflow, which is in turn related to the lack of explainability for most of the approaches based on modern algorithms. Indeed, it is a matter of fact that technological solutions based on neural networks and deep learning exploit preconceived Python routines downloaded from in-cloud repositories, whose content is partially or completely unknown even to the hard scientists that are coding the tools. As a consequence, we are currently experiencing a systematic lack of communication between wet scientists or medical doctors, who want to reasonably know at least the main methodological principles at the base of the data analysis approach, and hard scientists, who have tremendous difficulties in unveiling these principles for at least two reasons: because the mathematical language they need to use to explain the algorithms' strategy is too obscure for their clinical stakeholders, or, which is even more alarming, because even they are not fully aware of all the details of such strategies.

For all these reasons we think that the so-called 'algorithmic accountability' should be one of the main issues to address in order to reliably bring medical AI within the clinical workflow. An AIbased diagnostic/prognostic workflow is made of several complex and intertwined steps including the homogenization of the data at disposal, the mathematical design of the algorithm, the optimization of the model parameters, the implementation of the software tool, its validation, and EU regulations currently suffer a significant number of gaps as far as identification of responsabilities, complaint of insufficient transparency, taking charge of validation obligations are concerned. And, of course, this lack of reliability in the legislation implies that clinicians who are prone to use AI methods as a tool for supporting their decision are left in a particularly vulnerable position.

As a final remark we point out that a reliable, effective, unbiased realization of AI-based clinical workflows more and more urgently needs the design and construction of a completely renewed health care system that must rely on a deep, systematic, unprecedented integration of completely different disciplines. On the one hand, it is crucial that the mathematical literacy of biologists and medical doctors rapidly and significantly increases; on the other hand, biological and physical models must more and more be plugged into data-driven AI approaches in order to increase both their reliability and their predictive power. And, even more than this, we need to train novel interdisciplinary professional figures that are well acquainted with topics in both wet and hard sciences and that are able to navigate even in difficult and heterogeneous ethical and legal issues.

Psychotherapy and artificial intelligence

Linda Alfano; Rosagemma Ciliberti
Department of Health Sciences, University of Genoa, Genoa, Italy

Abstract

In recent decades, Artificial Intelligence (AI) has made significant progress and has been utilized in various fields, including the domain of mental health care. Specifically, the availability of digitally mediated psychotherapies has opened new perspectives in the field of mental health, introducing innovative elements that necessitate a careful reevaluation of traditional clinical tools and raise important ethical questions to be explored.

Introduction

Easier access to the virtual world and the constant increase in various forms of psychological distress have, in recent years, facilitated the emergence and development of numerous online psychotherapy platforms, more precisely termed as electronic psychotherapy or E-psychotherapy.

These digital platforms have overcome certain territorial and organizational obstacles that limited access to mental health services, offering individuals the opportunity to receive professional support from qualified psychotherapists conveniently from their homes.

The use of this medium became particularly evident during the pandemic, where there was a rapid digitalization of psychological support interventions. In this period, in fact, many people experienced increased anxiety and stress due to social isolation, economic uncertainty, and health concerns. International contributions have highlighted the fundamental role that the development of psychological support services played in the healthcare response to the COVID-19 pandemic (1).

As noted, the use of technology in the practice of psychotherapy is not a new concept. In fact, discussions about "telephone analysis" began in the United States as early as the mid-1900s. This approach was not only aimed at reducing costs and the inconvenience of travel or making up missed sessions but also at overcoming certain resistance to treatment (2, 3).

In the evolution of technology use, the internet has introduced a different approach, one that is highly adaptable and can be customized according to the ability to engage in real-time, face-to-face dialogue through a screen. This enables synchronous communication (which can simulate the therapy session and even the therapist's waiting room with the so-called "virtual reality"). Alternatively, it allows for other modes of communication such as chat and email, which are asynchronous in nature (4).

Recent applications of intelligent technologies in the field of psychological health represent a further and distinctly different area. The early, rudimentary contributions of computing to psychology date back to around the 1970s and were primarily focused on assisting therapists in symptom research and subsequent diagnosis formulation. The software provided a guide to the questions and areas to analyze, aiding in navigating the categorical diagnostic manuals, following a predetermined decision tree. In essence, it was a flowchart supporting the memory and expertise of diagnosticians, which generated valuable tools still widely used in clinical practice today.

Over the past few decades, the contributions of psychologists and neuroscientists have allowed for the testing of the first programs (chatbots) capable of responding to an interlocutor. These chatbots interacted with individuals, more or less aware that they were engaging with an automated software interface, providing listening and suggestions.

The second generation of artificial intelligence, introduced towards the late 1980s, led to a revolution in AI approaches, including the implementation of systems like chatGPT, a model based on neural networks that seeks to simulate the functioning of the cortical columns of neurons. These neural networks have basic rules and can autonomously learn from data without the need for explicit programming. This has paved the way for a range of innovative applications, such as image recognition, speech recognition, automatic translation, and much more. However, the creation of these neural networks with basic rules and the ability to learn autonomously (deep learning) has led to a progressive loss of control over the system, making it impossible to understand the thought process and motivations behind the provided responses. These new AIs are capable of passing the Turing test, meaning they can be mistaken for humans by an interlocutor who reads their messages without being able to see the sender. However, to a careful observer, it should not go unnoticed that the language used by these AIs is rather stereotypical, rigid, and, in a sense, quite similar to that of a neuroatypical individual.

Recently, Klos et al. evaluated the feasibility, acceptability, and impact of using the chatbot "Tess" to provide initial "psychological intervention" to Argentine students with social phobia and depression. According to the authors, the level of acceptance and the number of voluntary contacts with the chatbot showed positive feedback from users. Moreover, the reduction in symptoms compared to a control group demonstrated the clinical effectiveness of the tool. They concluded that a tool of this nature could help reduce the delay in which individuals with vulnerabilities seek help and enter treatment (5).

In another study, researchers tested the acceptability of a similar virtual interface, called "SimSensei", which even had sensors and hardware capable of detecting the interlocutor's facial expressions and gestures. The goal here was to increase the resemblance to human dialogue, incorporating feedback from non-verbal cues. The researchers highlighted the advantage of the sense of privacy and reduced shame on the part of the interlocutors when communicating symptoms and personal aspects of their psychopathology (6).

As of now, there are no official statistics available on the number of patients utilizing psychotherapy services through digital platforms, although the magnitude of the phenomenon is evident, as demonstrated by the continuous increase in the number of digital agencies providing this type of psychological support. The landscape is constantly evolving due to the rapid advancement of technologies, which not only impacts the ways individuals relate to one another but also the possibilities for intervention by various professional roles.
There are numerous potential benefits to the use of virtual communication technologies in psychotherapy. First and foremost, it addresses logistical, physical, temporal, and economic challenges, reaching individuals who may have difficulty seeking in-person professional help.
One of the primary applications of AI in psychotherapy is the automation of emotional support. AI-based virtual assistants can provide a safe and private environment for individuals looking to explore their feelings and issues. These virtual assistants can use natural language processing algorithms to understand the emotions of patients and respond empathetically and appropriately.
AI can also be used for continuous patient monitoring. Wearable devices and mobile apps can collect data on patients' emotional and physical states and send this information to psychotherapists. This monitoring can help identify early changes in mood or symptoms and allow for timely intervention.

AI can promote the practical implementation of the principle of equitable access to care, enabling people who traditionally had to forgo treatment due to work or geographical distance to benefit from these therapies.
The use of AI in psychotherapy can lead to more personalized treatment. Through data analysis, AI can more accurately identify the most effective treatment methods for each individual, taking into account their specific needs and preferences, thereby improving the effectiveness of the therapy. Among the various advantages, it's worth considering the promptness of access, as these systems allow patients to participate in virtual therapy sessions guided by virtual agents whenever they feel the need, without having to wait for fixed appointments.
A particularly sensitive target group for accessing online therapy consists of the so-called "IGeneration" (Generation of Networks). These are young people and adolescents aged between 16 and 25 who were born and raised in an environment entirely immersed in digitization and are generally very inclined to use technology.

These young individuals, growing up in the digital age and accustomed to managing most of their relationships and communications through digital devices such as smartphones, social networks, and messaging apps, have a profoundly different conception of psychotherapy compared to the traditional image of a patient lying on a couch and engaging in face-to-face dialogue with a therapist physically present in the same room.

This mode of interaction can be particularly beneficial for those dealing with specific issues such as social phobia, social withdrawal, or depression, as well as for members of ethnic minorities who may not have immediate access to support in their native language. Additionally, this tool has the potential to play a significant support role in monitoring emergencies and identifying situations of risk for patients promptly.

The Challenges: The Sense of Limitation
The use of AI in psychotherapy also presents significant ethical challenges that cannot be overlooked (7, 8). While this tool can be a useful and effective channel for timely intervention at the onset of symptoms, it should not be disregarded that such an approach should constitute the first step of a complete treatment process that goes beyond the digital realm.
Despite the evident advantages, chatbots cannot completely replace the importance of a human connection in therapy and emotional support. This connection requires the ability to fully understand the complex and varied nuances of human experiences.
Psychological therapy encompasses various components, including one that can be considered "psychoeducational." This component involves the act of explaining, naming subjective experiences, teaching problem-solving strategies, and more. The software interface appears well-suited to perform this task effectively and accessibly. However, it's essential to recognize that this dimension of therapy represents only a part of the overall therapeutic intervention. The more challenging and complex part of therapy goes beyond mere information transmission. It involves the ability to provide empathy, make sense of and give meaning to patients' subjective experiences, and coherently connect elements of their past with their current experiences, a practice that we could simplify by calling "interpretation".
In the United States, the government has already granted authorization to several insurance companies to recognize digital therapies as valid. Among these innovations, some forms of AI stand out, such as COCO and WISA, designed to address depression in young individuals. These systems are equipped with advanced algorithms that analyze the dialogue between the user and the chatbot, identifying patterns, words, and phrases that may indicate potential suicide risk. This ability to detect danger signals is based on word frequency, memory of previous situations, and data shared with the therapist.

In this context, AI demonstrates remarkable efficiency, even surpassing highly experienced psychotherapists. This is because the computing power of AI allows it to process vast amounts of data in an instant and to quickly discriminate situations at risk. Furthermore, the constant 24/7 access to digital therapies provides continuous and easily accessible support for patients. While some may hesitate to go to an emergency room during moments of crisis, they may feel more comfortable communicating or calling in such situations.

AI, therefore, serves a "frontline" or "emergency room" function, enabling the identification of emergency situations and promptly alerting mental health professionals to provide timely support in critical situations. However, despite the essential contribution of AI, the actual treatment must firmly remain within the therapist's domain.

The human element of therapy, such as empathy and the interpretation of the complex nuances of human experiences, the ability to establish deep connections with patients, represents a fundamental component of the therapeutic process. It has not only an irreplaceable ethical value but is also essential for therapeutic practice that software interfaces may support but not completely replicate at this time.

Providing a dynamic understanding of emotions, the complex psychological factors underlying a disorder or symptom, and a reading of the fundamental traits of personality and how it has developed, is a lofty goal that implies a subjective co-construction with the person in distress. The tools presented are more likely to represent important pieces of a journey guided by human sensitivity.

These are undoubtedly crucial elements within the therapeutic process. These aspects are based on an intricate interplay of interactions between the therapist and the patient, a dialogue that engages two human minds in a process of mutual understanding.

So, the collaboration between AI and human therapists can be extremely effective when it comes to recognizing danger signals and providing immediate responses. However, the human therapist remains the central element in offering a comprehensive and personalized treatment.

Furthermore, AI may have limitations in understanding the specific cultural and social contexts of patients, while human therapists bring a wide range of experiences and knowledge that can enrich the therapeutic process. Nevertheless, it is important to emphasize that the digital delivery of mental health services should not be automatically rejected or demonized. The significant opportunities that this technology offers for reaching a broader patient base, customizing treatments, and providing constant support call for a cautious approach. Such an approach should carefully consider the practical and clinical implications of this transformation to understand how to ethically and effectively integrate the possibilities offered by technology into clinical practice without compromising the quality of help provided to those in need.

Conclusion

The latest innovations in the mental health sector are adopting increasingly advanced technologies to make treatment more accessible, which, along with the benefits, still present risks and weaknesses. The primary limitation concerns the machine's difficulty in grasping the emotional, symbolic, relational, and anthropological dimensions of the data with which it must interact, limiting itself to interpreting only the empirical side of reality.

Furthermore, the absence of human supervision during interactions with the patient not only tends to make communication uniform and lacking in nuances but also carries the risk of dangerous self-diagnosis. This can lead to a harmful increase in self-diagnosis without the possibility of proper correction.

Moreover, the use of AI in the decision-making process could raise ethical concerns, such as patient privacy protection. The information collected during psychotherapy sessions, even if automated through

AI, must be treated with the utmost confidentiality and security. Patients' personal data, including details of their psychological experiences, represent sensitive information that requires strict protection.
Artificial Intelligence is changing the face of psychotherapy, offering unique opportunities to improve mental health and well-being. However, it is essential to address the ethical challenges associated with this transformation. Patient privacy, human supervision, responsibility, and accountability are just some of the issues that need to be carefully considered.
The key to an ethical and responsible integration of AI in psychotherapy is close collaboration between developers, mental health professionals, and patients themselves. Only through open dialogue and robust governance can we maximize the benefits of AI without compromising ethics and safety in mental health care.

References

1. Xiang, Y.T., Yang, Y., Li, W., Zhang, L., Zhang, Q., Cheung, T. & Ng, C.H. (2020). Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. *The Lancet Psychiatry*, *7*(3), 228-229.
2. Calderon J. (2004). Use of the telephone in psychotherapy. Edited by Joyce K. Aronson. Northvale, NJ: Jason Aronson, 2000, 468 pp., ISBN 0–7657–0268–1. International Journal of Applied Psychoanalytic Studies. 1. 201-205. 10.1002/aps.68.
3. Leffert, M. (2003). Analysis and Psychotherapy by Telephone: Twenty Years of Clinical Experience. *Journal of the American Psychoanalytic Association*, *51*(1), 101-130. https://doi.org/10.1177/00030651030510011301
4. Suler JR (2000). Psychotherapy in cyberspace: A 5-dimensional model of online and computermediated psychotherapy. *CyberPsychology and Behavior* 3, 2, 151-159.
5. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial. JMIR Form Res. 2021 Aug 12;5(8):e20678. doi: 10.2196/20678. PMID: 34092548; PMCID: PMC8391753.
6. Morency, L. & Stratou, Giota & DeVault, David & Hartholt, Arno & Lor-Lhommet, Margot & Lucas, Gale & Rizzo, Albert. (2015). SimSensei demonstration: A perceptive virtual human interviewer for healthcare applications.
7. Manríquez Roa, Tania & Biller-Andorno, Nikola & Trachsel, Manuel. (2021). The Ethics of Artificial Intelligence in Psychotherapy.
8. S. Carvalho, C. A. (2023). Ethical challenges of AI-based psychotherapy. The case of explainability. *Scenari*, (17). Retrieved from https://mimesisjournals.com/ojs/index.php/scenari/article/view/2598

Pros And Cons Of Open A.I.'s Chat GPT

Dr.Chinmay Shah
Professor & Head, Department of Physiology, Government Medical College, Bhavnagar, Gujarat, India. cjshah79@yahoo.co.in, Orcid ID : 0000-0002-4714-0129

Abstract: Publicly available generative AI (GenAI) tools are rapidly avilabe to all stakeholder. Due to emerging trade, popularity and abscent of national regulatory frameworks its use is not under any control. Thus compnnies are creating Generative AI i.e Chat GPT and man more and layman are using it without any fear. ChatGPT reached 100 Million monthly active users in January 2023, only one country has released regulation on generative AI in July. Every innovation are alwysed viewe positive side, after use people come to know dark side of any innovaiton. Similary Chat GPT or GenAI also has both side. In this chapter we have tried to cover pros and cons in regards to use of Chat GPT in healthcare and healthcare education to students.

1. Introduction :

Generative AI (GenAI) is an Artificial Intelligence (AI) technology that automatically generates content in response to prompts written in natural language conversational interfaces. While GenAI can produce new content, it cannot generate new ideas or solutions to real-world challenges, as it does not understand real-world objects or social relations that underpin language.

Moreover, despite its fluent and impressive output, GenAI cannot be trusted tobe accurate. Indeed, even the provider of ChatGPT acknowledges, 'While tools like ChatGPT can often generate answers that sound reasonable, they cannot be relied upon to be accurate[1].' Most often, the errors will go unnoticed unless the user has a solid knowledge of the topic in question.

By July 2023, some of the alternatives to ChatGPT includes Alpaca, Bard, Chatsonic, Ernie Hugging Chat, Jasper, Llama, Open Assistant, Tongyi Qianwen, YouChat, ChatPDF, Elicit, Perplexity, WebChatGPT, Compose AI, Wiseone and many more[2]. As of July 2023, the Image GenAI models that are available are Craiyon, E mini, DALL, E 2, DreamStudio, Fotor, NightCafe, Photosonic, Elai, GliaCloud, Pictory, Runway, Aiva, Boomy and many more[3]

Cloud computing is providing the computing capacity for the analysis of considerably large amounts of data, at higher speeds and lower costs compared with historic 'on premises' infrastructure of healthcare organisations. Indeed, we observe that many technology providers are increasingly seeking to partner with healthcare organisations to drive AI-driven medical innovation enabled by cloud computing and technology-related transformation[4-6]

As every coin has two sides, it hods true for ChatGPT too, lets look at two sides of Chat GPT (AI), first we will look on Pros and than Cons

2. Pro's

AI help in multiple way to improve our daily routine work, saves energy and time of worker. It can help is in multiple ways as mentioned below, we can categories it as help in health care delivery and in education[7]

Health care delivery

- **Administrative workflow: AI will help us in** doing paperwork and other administrative task, which can further freeing up employee time for other activities and giving them more face-to-face time with patients. i.e note taking, writing of summary report and staderadise discharge card as well as billing
- **Virtual nursing assistants:** AI **will** help us as virtual nursing assistant and provide answers to standardize question, forward reports to doctors or surgeons and help patients schedule a visit with a physician. These sorts of routine tasks can help take work off the hands of clinical staff, who can then spend more time directly on patient care, where human judgment and interaction matter most.
- **Dosage error reduction:** AI could be used to help identify errors in how a patient self-administers medications. It will definitely reduce Adverse events as well as near miss events
- **Fraud prevention:** Implementing AI can help recognize unusual or suspicious patterns in insurance claims, which will help society indirectly and reducing stress to real claimant
- **Improve the healthcare user experience :** AI could deliver more specific information about a patient's treatment options, allowing the healthcare provider to have more meaningful conversations with the patient for shared decision-making.
- **Increase efficiency in healthcare diagnoses :** According to Research, although it's early days for this use, using AI to make diagnoses may reduce treatment costs by up to 50% and improve health outcomes by 40%[8].
- **AI help for better health monitoring and preventive care:** As health and fitness monitors become more popular and more people use apps that track and analyze details about their health, they can share these real-time data sets with their doctors to monitor health issues and provide alerts in case of problems
- **AI can help connect disparate healthcare data :** One benefit the use of AI brings to health systems is making gathering and sharing information easier. AI can help providers keep track of patient data more efficiently.
- **AI help for governance in healthcare :** As AI becomes more important in healthcare delivery and more AI medical applications are developed, ethical and regulatory governance must be established. Issues that raise concern include the possibility of bias, lack of transparency, privacy concerns regarding data used for training AI models, and safety and liability issues.
- **AI works for the public's benefit** by Protecting autonomy, Promoting human safety and well-being, Ensuring transparency, Fostering accountability, Ensuring equity, Promoting tools that are responsive and sustainable

For Education There are several potential benefits of ChatGPT (Generative Pre-trained Transformer) for education, including (9):

- **Guidance and training**: Provide guidance and training to researchers, teachers and learners about GenAI tools to ensure that they understand the ethical issues such as biases in data labelling and algorithms, and that they comply with the appropriate regulations on data privacy and intellectual property.
- **Coach :** Generative AI can act as 1:1 coach for the self-paced acquisition of foundational skills, it should be reinvigorated and upgraded with GenAI technologies to foster learners' self-paced rehearsal of foundational skills. If guided by ethical and pedagogical principles, GenAI tools have the potential to become 1:1 coaches for such self-paced practice.
- **Project Based learning:** Generative AI also facilitate inquiry or project-based learning that aim to trigger higherorder thinking amongst learner. In simulated teaching conversation using AI can be used as it mimic human interaction.

- **Help Learner with Special Need:** Generative AI is of great help to learners with special needs like deaf and hard-of-hearing learners, and GenAI-generated audio description for visually impaired learners. AI can convert text to speech and speech to text to enable people with visual, hearing, or speech impairments to access content, ask questions, and communicate with their peers.
- **Personalised learning experience:** ChatGPT help to give Personalized learning experiences by analysing student's learning patterns and preferences and recommend specific learning resources that are tailored to their needs. When exams are around the corner, ChatGPT can help students prepare. It can recapitulate their class notes with emphasis on key terms. AI will Improving accessibility by use of chatbots and virtual assistants that can help students with disabilities or those who speak different languages to learn and participate in classroom activities. When exams are around the corner, ChatGPT can help students prepare. It can recapitulate their class notes with emphasis on key terms.
- **Help in Assessment:** AI will also help in Automated grading in grding essays and other written assignments automatically. This will save teachers a lot of time and provide students with immediate feedback on their work.
- **Assist teachers:** Using ChatGPT in higher education can assist professors in multiple ways like help in developing a comprehensive lesson plan for a course, creating questions like MCQs, true and false, fill in the blanks, etc., for a class test. AI also help to analyze students' assignments and aid teachers in grading and providing constructive feedback. It can help studnet and teacher by providing access to links containing additional educational resources for a course. AI will also provide us tips and tricks for increasing students' engagement and reducing troublesome behavior in the classroom.

Overall, we can definelty say that ChatGPT is versatile and it has become unabitable after it is upgarded to latest version . Chat GPT help to automate repetitive tasks, it help us to Optimizing a website for search engines (SEO). Time management can be definetly improved by use of ChatGPT. ChatGPT can help optimize some day to day work like Email drafting & Social media management. This pros of ChatGPT are because of its Speed, its tool as Natural Language Processing (NLP) and Machine Learning (ML), its efficiency and its use friendly experice

3. Cons

Despite the current public fervor over the great potential of AI, We recognise that there are significant challenges related to the wider adoption and deployment of AI into healthcare systems. Some of these cons are covered in this chapters, others go beyond the scope of this current book

- **Distorted Academic Integrity**: Academic integrity is the primary concern for using ChatGPT in higher education. Lack of Academic Integrity result in improper evaluation and ranking of learner, it will decrease students' abilities to brainstorm, think critically, and be creative with their answers.
- **Unaccurate information:** The information provided by ChatGPT can seem plausible and well-written, but it lacks insight and may not be necessarily accurate. It is generated from whatever is avilable on internate. It can be difficult to detect exactly which portions of the information are factually inaccurate. This will create wrong narrative and base for futhre information
- **Marginalise information:** Chatbots are trained on a massive dataset. If the dataset contains biases, chances are that some of the responses produced by Chat GPT will be biased. It may marginalise already marginalised information and prevent real information to reach to real world
- **Lack of emotional intellegence:** A human educator can understand the emotions of students and respond accordingly. ChatGPT, which lack EI and thus are unable to comprehend human emotions. Relying too heavily on Chat GPT for tasks such as writing and customer service can lead to a

decreased human touch and over-dependence on technology. It is important to balance the use of Chat GPT with human interaction to ensure a personalized experience for customers.

- **Hallucination:** Due to Ambiguities and inaccuracies in response, In language models, a phenomenon known as "hallucination" frequently occurs. Additionally, there are no references or citations for obtaining information. As such, it is not ideal to use this chatbot for research or electronic trailing alone.
- **Legal implication:** There's also a risk of misuse with AI-generated language models. The way they use internet information, they might respond in a biased or discriminatory way, which could upset others which may result in legal implicaitons. Chat GPT may have difficulty understanding the context of a conversation, leading to irrelevant or inaccurate responses. As with any AI tool, Chat GPT poses cybersecurity risks, especially when used for sensitive tasks like writing cybersecurity reports. There is a risk of confidential information being leaked, and hackers can use Chat GPT to generate convincing phishing emails.
- **Academic dishonesty:** If students use ChatGPT to generate written work without proper attribution or acknowledgement of its use, it could lead to plagiarism and academic dishonesty.
- **Unequal Access**: One of the significant challenges we face is the unequal access to AI tools for education. While some learners may have the financial resources to access advanced AI technologies, others may be left at a disadvantage.
- **Ethical Problem:** Working with AI raises concerns about fair compensation, working conditions, and the ethical treatment of individuals involved in tasks such as data labeling, annotation, and content creation tasks that require human intervention.
- **Fallcy in AI:** Fallcy in AI can be classified in to 1. Impossable Task as calimed by developer, it can be further devided in to conceptually impossible or practically impossible 2. Enginerering failure, it can be failure in desingm implementation or missing safety features. 3. Post deployment failure , which can be isuue in roburstness or issue in unanticipated interaction. 4. Communication Failures which may be misrepresentated capabilites or falcified capabilites[9]
- **Digital Poverty:** AI will leads to worsening digital poverty as GenAI relies upon huge amounts of data and massive computing power in addition to its iterative innovations in AI architectures and training methods but this data sets are not in reach of each ane very one which result in limited answer or reply to prompt

4. Reference:

1. Educator considerations for ChatGPT. San Francisco, OpenAI. Available at: https:// platform.openai.com/docs/chatgpt-education (Accessed 23 June 2023.)
2. Murphy Kelly, S. 2023. Microsoft is bringing ChatGPT technology to Word, Excel and Outlook. Atlanta, CNN. Available at: https://edition.cnn. com/2023/03/16/tech/openai-gpt-microsoft-365/ index.html (Accessed 25 August 2023.)
3. Guidance for generative AI in education and research. (2023, September 8). Unesco.org. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research (Accessed 25 October 2023.)
4. Satya Nadella announces strategic collaboration with Novartis. You Tube, 2019. www.youtube.com/watch?v=wMfsQE-D2q4
5. Lashinsky A. Tim Cook on how Apple champions the environment, education, and health care. Fortune, 2017.
6. Turea M. How the 'Big 4' tech companies are leading healthcare innovation. Healthcare Weekly, 2019.

7. IBM Education. (2023, July 11). The benefits of AI in healthcare. IBM Blog. https://www.ibm.com/blog/the-benefits-of-ai-in-healthcare/
8. ‘AI has already transformed medical innovation. Let’s not let fears of technology stop us from realising its incredible potential.’ (n.d.). Gsk.com. Retrieved November 9, 2023, from https://www.gsk.com/en-gb/behind-the-science-magazine/ai-medical-innovation-ethics-responsibility/
9. Inioluwa Deborah Raji*, I. Elizabeth Kumar*, Aaron Horowitz, and Andrew D. Selbst. 2022. The Fallacy of AI Functionality. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3531146.3533158

Bioethics Documents ready to be consulted

Bioethics Documents

Presidenza del Consiglio dei Ministri

CNB
COMITATO NAZIONALE PER LA BIOETICA

CNBBSV
COMITATO NAZIONALE PER LA BIOSICUREZZA LE BIOTECNOLOGIE E LE SCIENZE DELLA VITA

ARTIFICIAL INTELLIGENCE AND MEDICINE: ETHICAL ASPECTS

29 May 2020

CONTENTS

Presentation 3

1. Definitions of AI and recent developments 5

2. Developments in AI in the field of healthcare 7

3. Emerging ethical problems 9

3.1 AI in the doctor-patient relationship 9

3.2 The reliability of AI and the opacity of algorithms 10

3.3 AI and data: between privacy and data sharing 12

3.4 Consent and autonomy 13

3.5 Responsibility 14

3.6 Medical, technological and social training 15

4. Recommendations 16

Presentation

During a meeting between the President of the Council of Ministers and the Italian Committee for Bioethics (26 September 2019), a specific request was made by President Conte for a pronouncement by the Committee on the use of artificial intelligence (AI) in the field of healthcare. In the context of the speech addressed to the Italian Committee for Bioethics (ICB), President Conte underlined: "In the perspective in which we are heading, where technological innovation will further characterise our daily lives and where obviously, from a government perspective, we will push even harder for a decisive transformation in the digital sense, it is clear that artificial intelligence and robotics will play an even greater role and will challenge us to deal with moral dilemmas (...). I would therefore ask you to accompany me with your reflections on this at the very moment in which we are moving in that direction of development".

In order to provide a response to the query with reference to the applications in the field of health and medicine, the Italian Committee for Bioethics has set up a mixed group with the Italian Committee for Biosafety, Biotechnology and Life Sciences (ICBBSL) coordinated by Profs: Salvatore Amato, Carlo Casonato, Amedeo Cesta, Roberto Cingolani, Lorenzo d'Avack, Silvio Garattini, Laura Palazzani. The document was edited by Prof. Laura Palazzani.

The opinion, starting with a definition of AI, it analyzes its origins and most recent developments, with specific reference to the huge availability of data and computing power. The document highlights the opportunities and risks of AI and the main applications in medicine, including the context of the Covid-19 pandemic.

The Committees, in the context of the rapid evolution of these technologies, call on some elements of critical reflection for an understanding and evaluation of AI. As part of the doctor-patient relationship, they underline, on the one hand, the opportunities that can allow health professionals to reduce the time needed for bureaucratic, routine or dangerous activities, allowing them to be more available in the patient care relationship, on the other hand, they describe the risk that "automated cognitive assistance" could reduce the skills of doctors and healthcare workers. The document underlines the importance of tools that guarantee the reliability of AI, through validation, reducing, as far as possible, opacity, errors and possible discrimination due to technological and/or human causes. Given the enormous use of data, adequate protection of privacy is also essential, also considering the possibility of sharing data for "social good".

Informed consent remains an essential element of the doctor-patient relationship, despite certain difficulties, given by the informative process of the doctor and the not always simple and usual understanding of it by the patient. Particular attention is therefore also devoted to new training in the medical, technological and social fields. In this regard, the Committees believe it is essential to rethink the training of health professionals dynamically, with a flexible review of the study programs by interdisciplinary commissions, to combine the various competencies of AI in a transversal and interdisciplinary way and, at the same time, introduce the importance of ethics in the training courses of engineers, computer scientists, developers, with particular reference to ethics in the design and application of technologies. An important objective should also be the raising of public awareness within society regarding the opportunities and risks of new technologies, as well as a regulatory update on the profiles

concerning responsibility in the application of new technologies and the promotion of AI research in both the public and private sectors.

The identification of responsibility, from a legal point of view, requires an assessment of existing categories, given the pluralism of competencies between the designer, the software vendor, the owner, the user (the doctor) or third parties.

Auditions were organised with the internal experts of the two Committees: Dr. Amedeo Cesta, member of the Italian Committee for Bioethics (13 December 2019) and Prof. Roberto Cingolani, member of the Italian Committee for Biosafety, Biotechnology and Life Sciences (31 January 2020) as well as with invited external experts Dr. Alberto Tozzi, Head of Digital Medicine and Telemedicine Unit, Director of the Multifactorial Diseases and Complex Phenotypes Research Area at Bambino Gesù Paediatric Hospital (May 29, 2020); Prof. Carlos Romeo Casabona, Professor of Constitutional Law at the University of the Basque Country, Representative of Spain at the Council of Europe Bioethics Committee, Member of the European Group on Ethics in Science and New Technologies at the European Commission, Visiting Professor University of Rome Tor Vergata (13 December 2019).

The document was voted unanimously by the Italian Committee for Bioethics in the plenary session of 29 May 2020: Profs. Salvatore Amato, Luisella Battaglia, Carlo Caltagirone, Stefano Canestrari, Cinzia Caporale, Carlo Casonato, Francesco D'Agostino, Bruno Dallapiccola, Antonio Da Re, Lorenzo d'Avack, Mario De Curtis, Gianpaolo Donzelli, Silvio Garattini, Mariapia Garavaglia, Marianna Gensabella, Assunta Morresi, Laura Palazzani, Tamar Pitch, Lucio Romano, Massimo Sargiacomo, Monica Toraldo di Francia e Grazia Zuffa

Despite their not having the right to vote assent was given by: Dr. Maurizio Benato, the delegate for the President of the National Federation of MDs and Dentists Colleges, Dr. Carla Bernasconi, the delegate for the President of the National Federation of the Orders of Italian Veterinarians, Dr. Amedeo Cesta the delegate for the President of National Research Council, Dr. Paola Di Giulio the delegate for the President of the Superior Health Council, Prof. Carlo Petrini, the delegate for the President of the National Institute of Health.

Profs. Riccardo Di Segni, Luca Savarino and Lucetta Scaraffia absent from the session, subsequently assented.

It was also voted by the Italian Committee for Biosafety, Biotechnology and Life Sciences, on June 5, 2020, by Profs. Andrea Lenzi (President), Antonio Amoroso, Antonio Bergamaschi, Carlo Caltagirone, Roberto Cingolani, Fabio Fava, Paolo Gasparini, Maurizio Genuardi, Marco Gobbetti, Paola Grammatico, Mauro Magnani, Piero Angelo Morandini, Luigi Naldini, Ferdinando Nicoletti, Giuseppe Novelli, Pierfranco Pignatti, Roberta Siliquini, Paolo Visca.

1. Definitions of AI and recent developments

It is not straightforward to give a homogeneous definition of Artificial Intelligence (AI) especially in light of recent developments that have led to a widespread use of the term[1]. In general, this term identifies the sector of Information and Communication Technologies (ICT) which has the aim of imitating certain aspects of human intelligence with IT technologies, to develop "IT products or machines" able both to interact and learn from the external environment, and to make decisions with increasing degrees of autonomy.

Nevertheless, the term "intelligence", consistent with the English meaning of the word, does not designate strictly human qualities conferred on machines, but describes functions that make certain machine behaviours similar to those of human beings. The other absolutely particular aspect is that we are, for the first time, faced with technologies that systematically act as independent users of other technologies.

Developments in AI, following a period of stagnation, have gradually resumed during the last few decades[2] enabling to achieve its current significant impact. The great interest in AI has its explanations in technological progress, it has ensured that the results for years basically relegated to the field of scientific research, were attained first in industrial laboratories and from there have moved to the market.

To interpret the times we live in we must observe two aspects related to the evolution of technology: the computing power of computers, that has made possible the brilliant results of research which until only a few years ago only solved "toy instances"[3] of a problem, thanks to this power computers have now come to solve instances of real-life problems. The second aspect is linked to the increased availability of huge amounts of data[4] and information (which constitute the "examples" or basic elements for the machine), taken from ICT and the web, and to the development of algorithms[5] which have now become "executable": data and algorithms constitute the "learning" capability (machine learning), which allows the machine, on the basis of stored and archived information (data), to discover hidden relationships between data and the connection of information (algorithms)[6]. The algorithms integrate mathematical expressions, which find application in everyday problems to find associations, identify trends and identify

[1] See the Technical Report of the Joint Research Centre of the European Commission on *Defining Artificial Intelligence*: *Towards and Operational Definition and Taxonomy of AI,* 27 February 2020 (https://op.europa.eu/en/publication-detail/-/publication/6cc0f1b6-59dd-11ea-8b81-01aa75ed71a1/language-en/format-PDF/source-118380790).

[2] In classical AI one reasoned with a mathematics linked to philosophical logic, drawing deductions on the basis of causality. Today, however, the problem is that knowledge can be abstracted directly from data.

[3] One of the first was the chess game between the *Deep Blue* computer and the then world champion Kasparov.

[4] On the subject of data, see Italian Committee for Bioethics (ICB) *Information and communication technologies and big data: bioethical issues* (2016).

[5] The algorithm is a sequence of instructions that define the elementary operations that the machine must perform on the data to obtain the results. It is a systematic calculation procedure that solves a specific problem through a finite number of elementary instructions with a finite amount of data.

[6] Among other things, there are two distinct types of AI: a) weak emulation AI (*weak AI*) based on the principle that the essence of the functioning of the brain does not lie in its structure but in its performance; b) strong AI, simulation (*strong AI*) based on research to reproduce as closely as possible the physiology of the brain, possible extension of the weak one.

regularities within any set of data, at the basis of human behaviour, expressed by data and information[7].

The algorithms require a training phase starting from human-provided examples that form the basis from which to learn[8]. These two combined effects (data and algorithms) are the basis of current progress: while previously AI programs "were hand-trained" by the skill of their creators, they can now also "train themselves", with limits, and this has broken new ground. In the case of data available on the internet, automatic data acquisition can be programmed as they become available (some marketing or financial management tools use this mechanism).

Based on this "training" AI is able to predict, with varying degrees of probability. Correctness in the prediction of AI is proportional to the number and quality and accuracy of the data entered and the experiences stored on a given theme, but it could always fail should a case that has never happened before emerge. AI predictions depend on the data and algorithms with which the system is "trained": the predictions can be "wrong" due to the inaccuracy of the data provided or the use of unfounded assumptions. It is therefore necessary to take care of the "nourishment" or "training" of AI: in this context man maintains a central role. In this sense, it is not possible - to date - to speak of the "decision-making autonomy" of the machine.

This progress has affected areas of AI to different extents: many perceptual tasks (artificial vision, recognition of specific objects, interpretation of spoken and written language), to a lesser extent tasks similar to human reasoning, such as developing the ability to reason, understand intentions, elaborate arguments, as well as construct very articulated verbal speech. There is extensive discussion on the so-called "Logical capacity" of AI: what appears to be a logical-deductive process based on the typical concatenation of human reasoning (of causal association that starts from the description of reality and infers conclusions) performed by AI is, in truth, a dynamic model, based on very fast comparison and correlation with stored examples. AI, in this sense, applies a mathematical principle that highlights correlations between the data, but does not reason in a "logical" way in the proper sense. However, correlations are always probabilistic predictions and contain many limits and exceptions.

A particular area of *machine learning* is *deep learning* deriving from a process of imitation of the human brain, based on the creation of networks of artificial neurons. In *deep learning*, the machine extracts meanings by reasoning on large amounts of data: these are "automatic" (rather than "autonomous") learning methods, which, at the moment, generate "opaque" results that are not easily explained (the problem of the black box). These lead to superior performance results, which are sectorally close to human ones, albeit with some limitations.

---

[7] AI is not a new scientific-technological discovery but an area of research whose birth generally dates back to 1956 (year of the Dartmouth Summer Research Project on Artificial Intelligence to which the AI coinage certainly dates back), but which still dates back before the studies on the artificial neuron by J. McCULLOCH (*A Logical Calculus of the Ideas Immanent in Nervous Activity*, in "Bulletin of Mathematical Biophysics", 1943, 5 (4), pp. 115–133) and the studies by Alan Turing (*Computing Machinery and Intelligence,* in "Mind", 1950). The research area has gone through various periods: a first moment of vogue occurred in the eighties when there was a first marketing attempt (linked to the so-called "expert systems") which proved to be a limited and soon disappointing success. There followed a long period of slowdown in its advancements (also identified as "AI winter") basically dedicated to a return to the laboratories to resume basic studies, also strongly theoretical (A. VESPIGNANI, *L'algoritmo e l'oracolo*, Milano 2020).

[8] An equation in which a number of examples are accumulated to the left of the equal.

Much of the current fears and uncertainties concerning AI are based on the assumption of situations that are still unrealistic today, such as the "replacement" of the human decision-making capacity, or the "autonomisation" of machines that could escape human control[9]. Although theoretically possible, we are still far from this scenario. AI is a powerful tool, but it is an accessory to human decision. The problem today therefore is not so much that - feared by several parties – of misgivings regarding the "autonomy" of AI, but if anything, the problem concerns the fact that an expert system that becomes optimal in suggesting "decisions" to man risks reducing human attention with the possible consequence of reducing human skills (or deskilling)[10]. In this sense, it is important to reflect on the synergy between man and machine, and on the search for ways of intelligent "support" that allows man to have "significant human control" in terms of supervision and attention.

2. Developments in AI in the field of healthcare

2.1 Since the '70s, AI has been considered as an "emerging area" (called the area of expert systems) capable of carrying out reasoning on limited knowledge to imitate medical reasoning. Today in medicine there are already many AI applications aimed at improving healthcare practices.

Thus, AI can assist the professional in prevention and in classifying and stratifying the patient's conditions (reducing diagnostic uncertainty[11]); in understanding why and how patients develop diseases (reducing pathophysiological uncertainty); in considering which treatment will be most appropriate for them (reducing therapeutic uncertainty); in predicting whether they will recover with or without specific treatment (reducing prognostic uncertainty and increasing the prediction of the onset or evolution of pathologies), and so on.

In addition, efforts are currently being made to develop support for the doctor and health care workers[12] so that the most updated and appropriate guidelines

[9] Bear in mind that these concerns are found in several international declarations. For example, "the specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of rules of existing EU law meant to protect fundamental right" (European Commission, *White Paper On Artificial Intelligence - A European approach to excellence and trust*, 19.02.2020, p. 12). The UNESCO document on Robotics Ethics of 14, September, 2017 suggests a distinction between traditional deterministic computers and stochastic or probabilistic computers. The *Ethics Guidelines for Trustworthy AI*, developed by the High-Level Expert Group on Artificial Intelligence in April 2019 under the aegis of the European Union suggests the need to preserve the "*human-centric*" dimension of the new technologies. The European Group on Ethics in Science and New Technologies in the document of 20 March 2018 on *Artificial Intelligence, Robotics and 'Autonomous' Systems* also underlines the importance "that humans - and not computers and their algorithms - should ultimately remain in control, and thus be morally responsible".

[10] This is what happened in air transport: “intelligent” assistance to the pilot risks decreasing human attention and, to overcome this problem, the pilot is frequently trained on a simulator.

[11] The Dermosafe system, currently used in multiple hospitals, has a good chance of identifying critical situations and intervening promptly to stem the development of the tumour. And "*embodied*" artificial integration can give positive effects in the interaction with the patient in the prevention of degenerative diseases.

[12] In the text when referring to the doctor, where relevant, the indication also includes the reference to healthcare workers.

can be consulted, even while working on the ward[13]. Proper use of AI could also improve and make flows within healthcare facilities more efficient and smooth, from triage to emergency management or medical device selection, both in the pharmaceutical sector in relation, among other things, to the possible use of intelligent packaging for the packaging and sale of drugs. Furthermore, AI can be usefully employed in clinical trials and in the perspective of precision medicine.

2.2 As was the case with regard to the spread of Ebola, also in the Covid-19 pandemic which has hit the world since the end of 2019, AI is used to make a decisive contribution to the fight against SARS Co-V-2 virus. Since this is an epidemic and with a rapid spread on a global scale, it is essential to have equally fast tools that can be applied simultaneously in different areas. There are many uses made of AI and among these we indicate as prevalent: observation and prediction of the evolution of pandemic trends; diagnostic purposes of the pathology; search for a vaccine or a cure (AI can be of support to know the activity of drugs on metabolic chains and the structure of the receptors on which one wants to act); assistance to healthcare professionals and patients, by providing medicines and food and measuring vital signs; disinfection and decontamination; the sharing of knowledge and detection of misinformation; control and traceability of population behaviour[14].

2.3 In medicine AI, on the basis of data and analysis of data, exerts its effects in the interpretation of complex patterns. For example, in the interpretation of images, AI recognizes signals that the human eye can not distinguish. If AI has great experience of these signals, AI will perform well. There are many applications in the field of personalized medicine, in the analysis of big data in genomics, drug testing[15], in the operation of surgical robots (virtual reality, teleoperation, real-time image analysis). Robotic systems are integrated systems that refer to humans, integrating computational skills (thinking), data storage (memory), sensory systems (sight, hearing, touch) and actuators to generate physical actions (the skeletal muscle system). The synergy of these factors is applied inside the robot to operate according to the needs dictated by the operator (the doctor). The growth of AI in healthcare and in various sectors, in recent

[13] Currently, IBM is trying to make a version of Watson that is useful to the doctor; in particular, a system that provides the doctor, while he is operating, with specific information found in the literature of that area. Each month the Food and Drug Administration finds itself evaluating numerous diagnostic imaging algorithms. Also in the medical field, the FDA has proposed various guidelines and protocols to ensure *best practices* in managing these algorithms that evolve over time.

[14] See the Council of Europe document *AI and control of Covid-19 coronavirus*: https://www.coe.int/en/web/artificial-intelligence/ia-e-lotta-contro-il-coronavirus-covid-19. The initiative of the health facilities of Bergamo for new machinery is particularly significant, with the help of private funding, they have equipped themselves with robots to automate the procedures usually performed by operators, to prepare the saliva and mucus samples taken on swabs before being inserted into scanners that detect the presence of coronavirus. This automatic system is useful not only to lighten part of the lab technicians' work, which can be dedicated to other activities, but also to double the number of swabs analyzed per day.

[15] It is believed that it is possible to shorten the time spent on discovering a drug from 4, 5 years to one, with a cost cut of 80%. For example, Halicina is the first antibiotic discovered by an artificial intelligence algorithm. It is a broad-spectrum antibiotic that acts on difficult to treat bacteria, resistant to antibiotics. An anticancer, BPM31510, obtained with AI systems that sift through thousands of human tissues, has passed a phase 2 trial for patients with advanced pancreatic cancer.

decades, is also due to the growth of robotics in the civil field, used not only for domestic and recreational use (social robots), but also in the health-medical field for both diagnostic functions as well as for clinical practice in a strict sense. Also in this field there are multiple purposes: surgical support activities[16]; in the fields of diagnosis and prevention (with the aid of nanorobots); in the context of care, rehabilitation and personal assistance to the elderly, people with mobility problems, for people with autism[17].

3. Emerging ethical problems

The Committee notes the rapid evolution that is taking place in medicine and appreciates the enormous progress and extraordinary opportunities opened up by AI. The process has started and the transformation seems to be overwhelming, inevitable and irreversible.

In this evolving and transitional context, the Committee intends to recall some elements of ethical reflection, without exalting or hindering the development of technology, but rather to provide "critical" reflection for an understanding and evaluation of new technologies, in an attempt to understand how they really "work" and evaluate what is acquired in terms of potential. The goal is to identify the ethical conditions for a development of AI that does not forsake certain aspects of our humanity, in a new "digital humanism", for medicine "with" machines and not "of" machines. In the awareness that it is man who builds the technology and that technology is not a neutral tool, as it inevitably changes the doctor-patient relationship itself.

3.1 AI in the doctor-patient relationship

The impact on bioethical principles depends as much on the sectors of application of AI in the health field as on the identity of the doctor in light of this new support and on the overall function entrusted to the health service.

The use of intelligent machines and robots in medicine, insofar as they are and will be more efficient, precise, rapid and less expensive, seems desirable if we consider this replacement of man with reference to repetitive, boring, dangerous, demeaning or strenuous activities. If properly used, AI could reduce the time that professionals have to devote to merely routine bureaucratic activities, or activities which expose them to avoidable dangers, allowing them to have fewer risks and more time available for the patient.

Automation in the acquisition and interpretation of data, in the elaboration of diagnoses and in the identification of therapies or in the performing of the intervention itself cannot be completely independent of man, but requires constant verification, therefore it does not exclude the specificity of the relationship between doctor and patient. It is impossible to forget that each patient is sick "in his/her own way" and that personal contact is the essential element of every diagnosis and therapy. In this sense, the machine cannot replace the

[16] The robot-surgeon, equipped with forms of AI, carries out innovative support activities also allowing to perform the surgical operation by means of "virtual reality", with the aid of a pen connected to a tiny robot that directs the laser beam in the direction desired by the surgeon. Therefore, the doctor is immersed within the operating field and can carry out these activities remotely from the operating room.

[17] See a previous opinion of the ICB and ICBBSL, *Developments of Robotics and Roboethics,* (2017).

human being in a relationship that is built on the meeting of complementary areas of autonomy, competence and responsibility.

AI should be considered exclusively as an aid to the doctor's decisions, which remain controlled and supervised by man. It is for the doctor, in any case to make the final decision, as the machine solely and exclusively provides support for data collection and analysis, of a consultative nature. An "automated cognitive assistance" system in diagnostic and therapeutic activity is not an "autonomous decision-making system". It collects clinical and documentary data, compares them with statistics relating to similar patients, speeding up the analysis process of the doctor.

A problem arises: what happens when AI proves to perform better than the doctor? In some circumstances this is technically possible and this should be taken into account. It is in this specific space that the dreaded 'replacement' of man by machine could happen in the future.

But a further, more immediate consequence may be the delegating of decisions to technology. Delegating complex tasks to intelligent systems can lead to the loss of human and professional qualities. If the relationship of care is configured as a relationship of trust, as well as of care (Law 219/2017), the substantial role of the "human doctor" must be preserved as only the doctor possesses the skills of empathy and true understanding, which cannot be expressed by AI and it is precisely these skills alone which make such a relationship real. As suggested by some of the foremost experts on these issues four main components must be guaranteed: *Deep Phenotyping*, *Deep Learning*, *Deep Empathy* and *Connection*.

The predetermination of canons of behaviour and codes of conduct, such as protocols and guidelines, constitute support for the knowledge and experience of professional activity, but the requirements of diagnosis and care often oblige to go beyond predetermined models. It would be extremely serious if the space left to the supposed neutrality of machines led to the "neutralization" of the patient. The enormous potential offered by AI should be considered as a precious opportunity in which technique broadens the horizons of ethics, allowing to increase the patient's listening spaces and contact with the course of his/her illness. In this sense, AI would be a very useful tool that saves the time employed by the doctor in routine operations in order to gain more time for the relationship with the patient.

3.2 The reliability of AI and the opacity of algorithms

As mentioned, AI is made up of a series of algorithms: precise instructions and mathematical expressions to find associations, identify trends, extract dynamics from the data collected and entered. When the algorithms operate, they are considered ‘trustworthy and neutral' in themselves, only for the fact that their methods are represented through measurable, mathematical systems[18].

But it must be remembered that it is man (with the help of the machine) that collects and selects the data, and who builds the algorithms. In this sense, the AI system can be "opaque". "Opacity" refers to: the steps through which data are interpreted not being always explainable (transparent) and that they can also give

[18] As already highlighted, it is not always possible to reduce AI to synthetic logic such as that of an equation, precisely because AI is closer to a dynamic concept in which individual observations are compared to the atoms that make up the knowledge base.

discriminatory results. The discrimination does not come from the machine but from man who selects the data and develops the algorithms. This step implies a reflection on the "data ethics" supporting AI (which require both quality and interoperability) and on "algorithm ethics" (also called "algor-ethics"), which should be based on data that are not selected, alternatively on inclusive and non-discriminatory selections.

AI, even if it can reach a high degree of accuracy, it is not and cannot always be explainable. It is impossible for programmers and technicians themselves to explain how the system has achieved certain results (black box problem). Automation can lead to a lack of transparency on the logic followed by the machine: the machine does not provide, nor is it possible to trace, information on the correlations indicated or on the logic adopted to reach a conclusion or propose a decision (addressed to the doctor and/or patient).

The opacity surrounding the essential elements and the decision-making process by which an AI system can draw a conclusion, involve**s** the risk that health workers cannot validate and confirm, or reasonably discard, the proposal made by the system in an attempt to make their own decision. It is practically impossible for a human being to analyze the huge amount of calculations made by the algorithm and find out exactly how the machine managed to decide. This raises problems for the doctor in relation to the machine (whether or not to rely on the algorithms) and in relation to the patient, to whom the doctor cannot provide an explanation and transparent information.

Furthermore, algorithms, with the classification of people into groups or subgroups with profiles similar to those associated with certain schemes (clustering), may not take into account the variations that a particular patient may present. A care decision based exclusively on profiles elaborated on patients and in an automated way (through algorithms) can lead to the exclusion of treatment without offering in exchange an alternative, albeit presumed less effective, but nevertheless an indicated alternative. The risk arises, by classifying or stratifying patients into groups or subgroups on the basis of personal profiles obtained by them on the basis of various criteria or purposes, that discriminatory, stigmatizing or arbitrary decisions are made exclusively on the basis of these profiles or on the basis of considerations not related to healthcare (also indirectly linked, for example, to ethnic origin or gender). “Algorithmic discrimination" is possible, even in the medical field, with an impact on equity and inclusiveness. Inequities already exist in the health sector, but AI could accentuate and worsen them by creating and/or increasing the "gap" and inequalities. It is possible to avoid this drift with a broad and representative approach of useful data, continuously updated, for the development of algorithms.

Furthermore, it should not be forgotten that medical care also involves major economic interests[19], therefore AI can be oriented, through the construction of algorithms, to influence the doctor's decisions in various ways, for example by facilitating prescriptions through an increase or a decrease in normal values for a series of functional or biochemical parameters. Therefore, AI can bring to favour one class of drugs over another that has the same indications for a particular symptom or pathology. It can give preference to a diagnostic path which favours the use of certain reagents rather than others. It may suggest the use of certain more expensive equipment and technologies as an alternative to other cheaper

[19] One has only to recall that the drug market alone is worth at least 30 billion euros, with the addition of the market for diagnostic instruments, medical and rehabilitation devices.

ones. It can influence the doctor to prescribe treatments rather than stimulate the patient to improve good lifestyles[20].

These, among other risk reasons, push the Committee to believe that accurate controls must be made, also through the validation of the algorithms, in order to obtain the most probable certainty that the introduction of various forms of AI are beneficial to improve the quality of the services of the National Health Service. In other words, all the "products" of AI must be compared, through studies conducted with the rules of controlled clinical trials, with decisions that are made independently of AI by groups of competent doctors[21]. Without prejudice to the fact that controlled clinical studies remain the "gold standard" for the demonstration of the efficacy and safety of treatments, it must be borne in mind that when we talk about the application of AI in medicine, it refers to software[22]. With the problem that the mechanism changes over time and validation requires monitoring and further checks.

Only if it emerges from these studies that AI has a better performance than that of doctors, should it be accepted and used. However, the meaning of "performance" must be considered: for example, the segmentation of diagnostic images can be done with high level of quality by a doctor, but an AI system takes a fraction of time to perform the same operation and never tires.

This is particularly important for the improvement of the quality of the services of the National Health Service in the interests of citizens. It is also essential to set up facilities suitable for public interest research to take charge of the development of AI through public funds. It will therefore be necessary to demonstrate AI safety on the basis of control starting from the data base, the advantage in terms of benefits and risks, in a clinical sense, and cost-effectiveness, as well as the diffusion and sustainability of these technologies throughout the territory and over time. Only in this way will it be possible to demonstrate the reliability of these systems through certifications that guarantee their usability in clinical practice. Only in this way can there be the entrusting of complex tasks in order to support the trust relationship between patients and AI.

3.3 AI and data: between privacy and data sharing

In medicine, AI "feeds" on data: data is indispensable for the "training" of the machine and are the basic elements of the construction of algorithms, mathematical models that interpret the data. The availability of data (clinical data, images, genetic data, etc.), the accuracy and quality of the data, the interoperability of the data (through standardization and classification criteria) are the necessary conditions for the developments and applications of AI. Since every AI system is based on data, the problem of verifying, selecting, preparing and supervising data from human beings emerges, avoiding the errors of data

[20] See R. SPARROW, J. HATHERLEY, *High Hopes for "Deep Medicine"? AI, Economics and the Future of Care,* in "The Hastings Centre Report", 18 February 2020.

[21] For example, if a *learning machine* is programmed to diagnose and treat lung disease, it must be evaluated against the decisions of a group of doctors with specializations in pneumology.

[22] It is the same theme of *digital therapeutics* that is prompting towards some reflection (real world evidence) which also have regulatory implications. Beyond the methodological question, there is still uncertainty as to which regulatory body will have to deal with the issue. A discussion is underway both at European Medicines Agency (EMA) but also at Italian Medicines Agency (AIFA).

collection and classification, also providing AI mechanisms for checking and verifying correctness.

The insistence on the protection of privacy and confidentiality is often pointed out as being an obstacle to the development of AI. Those who intend to apply AI insist on the need to dispose of data in a broad field of action, on a global level (therefore also with transfer of data to other countries) and storage of data, but also storage of samples and associated images, over time. Data not fully anonymous but pseudonymised that allow traceability, identification in cases of importance of communication of the results, with appropriate conditions to prevent improper disclosures.

The huge collection of data, necessary for AI, also highlights the risk, related to the use of data and the crossing of data, of both intentional and accidental re-identification, raising the problem of privacy, which in this context tends to "vanish"[23]. To the point that it is believed that technologies are becoming increasingly "opaque" and the users "transparent".

In the AI era and the need for the use of data for medical research, questions arise with regard to the possibility of "sharing" data (*data sharing*) as a "social good" for the advancement of scientific knowledge.

There are methods and technologies for performing data transactions while preserving data security (one of the technologies is the family of *block-chain* applications).

This sharing is, in any case, guaranteed by the exclusive use for research purposes, which enables a return of information and sharing of clinically relevant results (*benefit sharing*).

There is wide debate, even on a regulatory level, of the applicability of the General Data Protection Regulation (GDPR) to AI scenarios, where it is unrealistic to protect privacy and guarantee data control, in the global research area (ICT) and in times that cannot be defined a priori.

3.4 Consent and autonomy

The favourable efficacy/risk ratio and the requirement of informed consent and autonomy are fundamental rules in the doctor-patient/healthcare worker relationship, as they protect the right to life, health, dignity of the person, self-determination.

It follows that this must be central even if the doctor wishes to make use of the data collected by AI and the development of robotics in the healthcare treatment. But the informative process is far from easy to implement and autonomy/consent is complicated by AI which arouses a sense of disorientation given the speed with which technologies are radically changing the known world. It is not easy for the patient-person to imagine the consequences that could arise from these new technological advances: It is the doctor who must act as mediator in this communication. Complex terminologies, words that may sound

[23] On the subject of privacy, the ICB intervened in the document *Information and communication technologies and big data: bioethical issues* (2016) underlining that as part of the "data processing when requesting information, it must always be accompanied by an explicit informed consent", in a transparent, complete and simple way, specifying "who collects and who will use the data, what data, how it is collected, where it will be stored and for how long, for what reason and for what purpose", specifying revocability. In the opinion *Mobile-health apps: bioethical aspects* (2015), the ICB expresses awareness "of the difficulty of achieving an informed consent and of protecting the privacy of users in this new field of application".

mysterious, are found in the new healthcare procedures (*machine learning, deep learning, neural networks, big data, algorithms, cloud, etc.*) making the consent to new healthcare treatments increasingly complex and given perhaps more through trusting the doctor than on actual understanding. Informed consent to AI-based healthcare treatments may impact patient autonomy. Certainly the patient sees the traditional relationship of alliance with the doctor change: still having very confused ideas about the applications of AI, the patient appreciates its advantages, but does not fully understand its risks.

It is therefore an ethical and legal obligation that those who undergo such innovative health treatments, through AI, are informed in the most appropriate and comprehensible way for the patient as to what is happening, to be (if necessary) subject to experimentation and validation; to be aware that what is applied to them (on a diagnostic and therapeutic level) implies advantages, but also risks. It must be explicitly specified in the informed consent if the applied treatments (be they diagnostic or therapeutic) are only from a machine (AI, robot) or if and what the areas and limits are in human supervision or control over the machine. These difficulties in the providing of understandable and exhaustive information, given by the doctor to the patient (difficulties regarding both the doctor's communication and the patient's reception) when employing treatments that make use of AI are augmented by the opacity of the algorithms.

3.5 Responsibility

Automation in medicine can contribute to the reduction of accidents and mortality (increase attention and the accuracy of the doctor's actions, enhancing its use even in routine procedures, etc.), but, as mentioned above, it is not without its risks. Machines can be poorly planned and poorly employed. Therefore, the issue of liability is one of the most delicate and complex problems that arise with the use and development of new AI systems. In particular, the problem receives increasing attention in terms of policy and legislative strategy. The attempt is above all to clarify whether accountability for certain decisions made through an intelligent system should be attributed to the designer, the software vendor, the owner, the user (the doctor) or third parties. The possible occurrence of accidents should be traced and analyzed as is the case for any medical error.

Any evolution that in the medical field ends up changing the doctor-patient relationship, should provide for an intervention of the law that governs innovations in accordance with the existing "system", thus creating guarantees both for the patient and the covering of new risks, and at the same time for the work of the doctor. The factor that poses new requests for legal mediation is not so much the presence of an intelligence with the possibility of self-learning, but the fact that AI has an "author" who creates it and who may not coincide with the "producer" of the good that incorporates it, the "seller" and the user and for whom the problem arises of outlining rights, limits and responsibilities. In many cases, command of action remains with the doctor, who is however not directly the agent of the action, being in fact sometimes in a different distance place. The fact that there is a "chain of command" to which the responsibility of the action can be traced may suggest that the action is less subject to chance and improvisation, but each link in the chain has its fragile point and given the complexity of the gestational structure and of the action it is not obvious to say who in the end is responsible for what eventually happens to the detriment of the patient or if this responsibility is unique. These autonomous and distinct responsibilities should be able to be

directly asserted by the end user of that product, the patient, not only through the traditional contractual plan that binds the doctor and the healthcare facility. In addition to these responsibilities there are also those differentiated in the context of the relationships between professionals (designer, validator, software vendor, programmer, etc.), who have contributed to the formation of the doctor-patient chain, without there being a prior proposal, a prior formal act, so that it is extremely difficult to recognise a contract equally. We are in a logic that brings us closer to the category, developed by jurisprudence, of "social contact responsibility", which hypothesises an obligation that is linked to the duty of diligence in observing the rules of art that it professes.

These are all aspects in which the analysis of new responsibilities and conceivable new rules, as well as new evolutionary interpretations of existing rules, make collaboration between legal and medical sciences indispensable, since the former have to deal with the latter and vice versa. An interdisciplinary continuous reflection that sees the two competencies "talking to one another" is both opportune and indeed necessary, also in order to outline the future structure of possible multiple medical responsibilities connected with AI.

3.6 Medical, technological and social training

Today, the medical world and healthcare professionals are not fully trained, with few exceptions, to use the results of AI research responsibly. It is therefore very important to act on two fronts: on the one hand to insert the problems deriving from AI in the activities of Continuing Medical Education (ECM) carried out independently and on the other to undertake a reform of Medical Schools, as well as the schools of healthcare workers.

The inclusion of AI in the education of doctors and health professionals[24] falls under the so-called *reskilling* of employees, i.e. reconverting workers (in this case healthcare workers) in the face of developments in emerging technologies. However, training salaried healthcare workers to occupy the same positions, but which imply new needs and professionalism, will be more complicated and expensive than creating new jobs for people already trained in understanding AI in the medical field[25]. This gives rise to the concern of the European Group on Ethics in Science and New Technologies (EGE), in the opinion *Future of Work, Future of Society* 2018 "*skills polarization*" that can hide new forms of discrimination, excluding those who are unable to learn the new required "skills". The problem of new professions, even in the medical field, remains therefore that high-level skills will be required. This discussion falls within the field of new diagnoses and therapies; continuous updating is essential for doctors and healthcare professionals.

The other path is the reformulation of medical education programs, allocating a significant part of the training of future doctors to the problems deriving from the digitalisation of medicine which is the basis of the AI technologies that future doctors will have to take advantage of, being able to understand its advantages, limits and dangers. The institutionalisation of interdisciplinary courses for the

[24] In recent years Europe seems to have become increasingly aware of the importance of the problem, just as several committees are beginning to take an interest in it, outlining the orientations of ethical and legal reflection in the field of AI.

[25] It follows that despite the emergence of many new professions, we may be witnesses to the formation, as already suggested by several parties, of a class of unemployed and "useless" individuals.

training of health professionals to a constant adaptation to technological change and to the possible "convergence" and transversality of traditional disciplinary sectors (e.g. medicine and computer science or physics or data science with foundations of computer science and AI, components of clinical ethics, bioethics and biolaw) is desirable.

Training must also be renewed in the field of technology, introducing ethics and bioethics training courses for engineers, computer technicians, computer scientists, with particular reference to ethics in the design of technologies (*ethics by design/in design/for designers*) and in the planning, methodology and application of technologies. This is the only way to ensure the ethical awareness of those building the technologies, in order to allow principles and values to be present from the beginning of the technological design.

It is also desirable to promote public debate on the developments and limits of AI in medicine, so that all citizens can acquire the basics of "*AI literacy*", active participation in social discussion. In the long term, it is hoped that the introduction of science as an essential part of culture in schools can lay the foundations for an understanding of the presence of AI within various sectors. These are the prerequisites for a possible overcoming of the "digital divide" in medicine, promoting greater inclusiveness.

4. Recommendations

In the light of the previous analysis, the ICB intends to recall some ethical principles of reference in the context of the use of AI in medicine. In the face of the progress in this area considered as "transformative" and "disruptive", especially in the field of health protection, the Committee intends to promote ethical reflection in balancing the human dimension and the artificial dimension, without mutual exclusion. In the belief that: exclusion of the artificial takes away many opportunities for man; exclusion of the human raises many critical issues given the limits of the artificial. We must avoid excessive hopes, but also excessive fears, adopting an attitude of trust and caution.

Committee recommendations:

- prepare *ex ante* accurate controls for the "training" of machines on the basis of quality data, that are updated and interoperable and conduct adequate experiments in the context of AI to guarantee safety and efficacy in the use of these new technologies as well as encouraging research in technology validation and certification tools and surveillance and monitoring, as indispensable elements for creating a "social pact of trust and reliability" of technologies in the medical field; it would be advisable to integrate the figure of a computer scientist or an AI expert into ethics committees for experimentation, and also update the legislation on experimentation with reference to software in the medical field;
- in the context of the doctor-patient relationship, informing patients in the correct way, especially during this transition period, regarding the risks and benefits of using AI with reference to specific applications (and also of the limits of explainability of "opaque" technologies), in order to ensure full awareness of the choices and also assuring alternative paths to the extent that resistance to accept the new technologies emerges; guarantee, in the applications of AI for health, a broad and representative (non-selective and discriminating) approach and an area of "significant human control" of human-machine interaction and collaboration, to protect overall correctness and patient-doctor communication as a field of care;

- rethink the training of health professionals in a dynamic way with a flexible review of the study programs by interdisciplinary commissions, for constant adaptation to technological change, also thinking about the possible "convergence" of paths in traditional disciplinary sectors (e.g. in the faculty of medicine, medicine and computer science or physics or data science and symmetrically, in the faculty of law/human sciences with fundamentals of computer science and AI);
- introduce the importance of the ethical principles of autonomy, responsibility, transparency, justice in the codes of conduct and the training courses of engineers, computer scientists, developers, with particular reference to ethics in the design of technologies (*ethics by design/in design/for designers*), ensuring technology that is oriented towards incorporating values and ensuring the centrality of the patient;
- create public awareness in society regarding the opportunities and risks of new technologies, so that citizens can participate critically in the debate on AI, without blind trust and not even an excess of concern, being aware of the choices and implications of digital healthcare: such promotion can also take place through organising conferences for schools and meetings with citizens, which the ICB regularly proposes;
- request, on a regulatory level, an update on the profiles concerning responsibility in the application of new technologies;
- promote research on AI, not only in the private sector, but also and above all in the public sphere of the National Health Service (NHS).

Mid-Term Review

of DPPA Strategic Plan 2020-2022

25 August 2021

Dr Ian Wadley, Independent Consultant

DISCLAIMER:

The evaluation report reflects the personal views of the author and does not necessarily represent the policies or position of the Department of Political and Peacebuilding Affairs.

Table of Contents

EXECUTIVE SUMMARY, KEY FINDINGS & RECOMMENDATIONS 2

1. OBJECTIVES AND SCOPE OF THIS REVIEW 8

2. IMPLEMENTATION OF THE 2020-2022 STRATEGIC PLAN TO DATE 9

2.1 A CHANGING OPERATIONAL CONTEXT FOR DPPA 9
Geopolitical and conflict trends 9
Constraints on UN Resourcing 9
Reform of the UN peace and security pillar 10
COVID-19 pandemic 10
2.2 DPPA'S RESPONSE TO THE COVID-19 PANDEMIC 10
2.3 IMPLEMENTATION OF THE STRATEGIC PLAN'S RISK-RESPONSE MODEL 13
2.4 IMPLEMENTATION OF DPPA'S STRATEGIC OBJECTIVES 13

3. DPPA'S STRATEGIC PLANNING TOOLS IN PRACTICE 14

3.1 THE STRATEGIC PLAN 2020-2022 IN PRACTICE 14
The strategic logic of DPPA (Theory of Change) 15
Aligning the strategic logic of DPPA Divisions 16
Aligning DPPA & DPO strategy 17
Risk management and the Strategic Plan 2020-2021 17
3.2 THE DPPA RESULTS FRAMEWORK IN PRACTICE 19
Scope for a dashboard view of performance 20
Improving on the relevance of indicators 21
*Highlighting documented peace outcomes ..... **Error! Bookmark not defined.***
Highlighting hidden results 22
3.3 THE DPPA ANNUAL WORKPLANS IN PRACTICE 23
3.4 OVERVIEW OF DPPA RESOURCES AND THEIR ALLOCATION IN PRACTICE 25
DPPA Resources 27
DPPA Reporting on resource allocation 28

4. CONCLUSION 33

ANNEX: METHODOLOGY 34

Executive Summary, Key Findings & Recommendations

Purpose and scope of this Mid-Term Review

The UN Department of Political and Peacebuilding Affairs (DPPA) commissioned an independent Mid-Term Review to assess: i) the progress made by DPPA in the first 17 months of its 2020-2022 Strategic Plan, from 1 January 2020 to 31 May 2021, and ii) how well DPPA's strategic planning tools have served the Department to date.

This is DPPA's first Strategic Plan following the restructuring of the peace and security pillar. Conducting a Mid-Term Review half-way through the implementation period enables the Department to gain a view of progress to date, and to identify scope for adaptation in the second half of the strategy period. The review was also requested to include concrete recommendations for improvements to both implementation and results reporting, while factoring in changes in operating context due to the COVID-19 pandemic.

The primary audience for the report is the Under-Secretary-General for Political and Peacebuilding Affairs, Rosemary A. DiCarlo, along with senior directors and staff within DPPA. The findings and recommendations of the Mid-Term Review are intended to serve as a basis for decision-making by DPPA leadership, in order to improve DPPA's performance and impact as outlined in the Strategic Plan. In keeping with DPPA's commitment to inclusivity, accountability and transparency, this report has been written with a broader audience in mind, including stakeholders, donors, counterpart organisations and researchers.

A changing operational context for DPPA

DPPA's 2020-2022 Strategic Plan covers a period of increasing tension in international affairs, with the multilateral system under pressure.

The instability caused by armed conflicts, terrorism, natural disasters, displaced populations, political crises, increasing technological disruption, along with climate and environmental change, were exacerbated by the unchecked spread of the COVID-19 virus and the ongoing global pandemic in 2020-2021. This has multiplied risks and inequalities for vulnerable populations in conflict situations, notably for women and girls, children (especially unaccompanied children), detainees, refugees or displaced persons, those who have disabilities, the elderly, and people who belong to a vulnerable minority group.

The COVID-19 pandemic global outbreak in 2020 also coincided with constraints on the UN's funding position due to delays in regular budget contributions by some Member States, which prevented the Department filling a number of critical vacant positions. Despite these constraints, DPPA continued to advance its conflict prevention, peacemaking and peacebuilding work, thanks to Member States' voluntary contributions to DPPA's Multi-Year Appeal fund (MYA), which received $35.9 million of the $40 million requested in 2020.

Another major feature of DPPA's operational context in 2020-2021 was determined by the reforms to the peace and security pillar that entered into force on 1 January 2019.

DPPA's response to the COVID-19 pandemic

DPPA's rapid and flexible risk-response model provided a valuable means to manage the threat to global peace and security posed by the COVID-19 pandemic. Identifying the potential for COVID-19 to exacerbate conflict patterns, DPPA supported the analysis and response of over 30 Special Political Missions (SPMs) and 100+ UN Country Teams, assessing the impact of COVID-19 on conflict dynamics, and proposing actions to foster peacemaking and prevent violence in the context of the pandemic.

DPPA's response to COVID-19 included the rapid and flexible work of the Peacebuilding Fund (PBF), which worked with UN Resident Coordinators and partners to make adjustments during 2020-2021 to nearly half of all ongoing PBF-funded programmes, mitigating risks of violent conflict posed by the pandemic, including countering hate speech and disinformation, addressing social cohesion, and helping to ensure equitable access to health care.

DPPA recognised the impact of COVID-19 on the most at-risk members of conflict-affected populations, particularly women and girls. DPPA continued to push for implementation of the Women, Peace and Security Agenda (WPS), including through gender responsive analysis, targeted efforts to support women's meaningful participation and to address conflict related sexual violence, and dedicated funding for gender programming.

The Secretary-General's appeal for a global ceasefire on 23 March 2020, supported by DPPA efforts, received endorsement from over 180 Member States, as well as a broad range of regional and civil society organisations. However, a lack of tangible support from actors with influence over conflict parties prevented the ceasefire reaching its full potential.

DPPA adapted its working methods in response to the COVID-19 pandemic, reallocating resources and putting into practice pre-existing efforts promoting the safe use of digital technologies for conflict prevention and resolution. The Department encouraged innovative approaches across its work, including shifting meetings of the Security Council and its subsidiary organs from solely in-person to fully virtual format. DPPA also pivoted to offer many of its training and learning opportunities online.

DPPA and its SPMs designed and implemented new hybrid models of mediation, combining in-person and digital interactions, as well as other tools such as digital focus groups powered by Artificial-Intelligence software. Some of these new virtual practices are likely to be continued even after the pandemic subsides, while other aspects of DPPA's mandate are likely to depend on in-person engagement with counterparts and cannot be readily moved to a purely virtual format.

Implementation at the mid-term of the 2020-2022 Strategic Plan

DPPA is performing soundly in the implementation of its strategic goals, based on the evidence available for this Mid-Term Review of the 2020-2022 Strategy.

The Department reported that it met or exceeded more than 79 per cent of its own performance measures under the Strategic Plan for 2020, and is on track to deliver similar performance in 2021.

In the first half of the strategy period, DPPA launched flexible and timely risk-responsive initiatives, including a total of 188 deployments of teams or individual experts in response to requests.

DPPA's Strategic Plan in practice

Finding : Evidence reviewed for this Mid-Term Review indicates that the Department's 2020-2022 Strategic Plan is useful in aligning effort, clarifying strategic logic through the Department's 'Theory of Change', and communicating the value of DPPA's work to outside audiences.

Recommendation 1:

⇒ **DPPA should consider developing a more operationally-focussed Theory of Change for its next strategy cycle, better reflecting the process through which DPPA identifies risks of conflict, reaches relevant actors and networks, engages them in dialogue, and exerts influence for peace. Additional benefit might be obtained if each Division were to formulate a theory of change or strategic logic as part of its workplan, supporting the whole-Department strategic logic.**

Finding : The Secretary-General's 2019 reform of the UN peace and security pillar means that DPPA and DPO are now aligned by a common vision statement, collaboration between Executive leadership, a shared set of objectives, a shared regional structure, and an emerging shared risk management approach, while retaining their distinct mandates and responsibilities.

Finding : The bulk of attention in DPPA's May 2020 risk register is devoted to the COVID-19 pandemic, for good reason. In most cases the risk definitions are very broad and incorporate multiple risk factors into a single risk, which is likely to make the assessment of gravity and likelihood more complex for DPPA staff and management.

Recommendation 2:

⇒ **Noting the approval of the UN Secretariat-wide risk register in July 2020, and bearing in mind DPPA's ongoing work to assess and manage risks on an organisation-wide basis and in conjunction with DPO, additional benefit might be gained by DPPA**

systematically covering additional categories of organisational risk, and directing attention to the most significant identified risk factors.

DPPA's strategic planning and reporting tools in practice

Finding : DPPA's robust and functioning Results Framework supports the Department's requirement for transparency, accountability, and a strong value-for-money claim under the 2020-2022 Strategic Plan.

Recommendation 3:

⇒ **DPPA should consider providing a one-page 'dashboard' view of its performance against strategic goals, combining both qualitative and quantitative assessments, while avoiding the temptation to reduce all of the Department's work to mere numbers.**

Recommendation 4:

⇒ **DPPA should improve on the relevance of indicators where feasible, focusing attention on DPPA's impact on the ground.**

Finding : DPPA does not fully report on significant but hidden interim results such as the cultivation of trusted networks with key actors, and the quiet but substantial support provided by DPPA to UN Resident Coordinators, Country Teams, Development Coordination Offices, SPMs, and other partner organisations. Senior DPPA staff advised that these kinds of achievements are routine activities rather than results, and advised against seeking to quantify these aspects of DPPA's work.

Recommendation 5:

⇒ **Given the significance of interim results such as trusted access and engagement with the right actors, DPPA should examine whether it might be possible to report on the value of these hidden achievements in an aggregated and de-identified manner, without jeopardising peace operations.**

Finding : The annual DPPA work plan process is rarely used for adaptation midyear in response to changing contexts, although it has potential to serve this purpose if treated as a 'living document' owned by the divisions. There is additional scope for DPPA divisions to use the workplan for adjusting priorities and planned activities during implementation, rather than seeing the workplan process primarily as a reporting tool for donors. Despite the admirably concise nature envisaged for the workplan reporting template, DPPA faces the ubiquitous risk that the focus of operational reporting drifts towards reciting generic activities, rather than specific valued results.

Recommendation 6:

⇒ **To counter the natural tendency towards activity reporting, DPPA might consider requiring divisions to introduce each section of their reports with a headline statement and two-line summary, to draw attention to the most valuable result generated, or the most significant risks identified and managed, and why this work was significant.**

Overview of DPPA resources and their allocation in practice

Finding : It is difficult to obtain a summary of the financial resources available to DPPA in its peacemaking, peacebuilding and conflict prevention mission. DPPA's extra-budgetary resources are obtained by the Department under the MYA, while its regular budget allocation from the UN General Assembly is managed separately. An overview of the various regular budget and extra-budgetary funds, and the total amount available to the Department would be a useful addition to DPPA's reporting, if this is feasible.

Recommendation 7:

⇒ **DPPA should consider whether it is possible to provide a summary overview of the annual financial resources available to DPPA in its peacemaking, peacebuilding and conflict prevention mission, including regular budget and extra-budgetary funding, and noting those resources that fall outside DPPA's Strategic Plan.**

Finding : DPPA's allocation of funds under the MYA is aligned with the Strategic Plan.

Recommendation 8:

⇒ **DPPA should focus financial reporting under the MYA on value creation and 'return on investment', rather than the rate of expenditure of allocated funds.**

Finding : DPPA's reporting obligations are multi-layered. In addition to its reporting under the Strategic Plan 2020-2022, the Department's core budget is subject to reporting under the UN regular budget framework approved by the UN General Assembly, which is not directly connected to the objectives set out in the Strategic Plan 2020-2022. Each year around $700 million is allocated from the UN regular budget to SPMs managed by DPPA, along with $11 million in extra-budgetary funds. DPPA supports, guides, and oversees SPMs, but it does not report directly on their resources and results in the Strategic Plan Results Framework or in DPPA's own Annual Reporting. This means that the value-for-money offered by the SPMs managed by DPPA is not contained in a single report. DPPA's public MYA reporting already provides relatively detailed examples of selected SPM results in narrative form, but does not feature a summary cost/benefit overview. The SPMs report

separately to the UN General Assembly (albeit in a format primarily focussed on results based budgeting and planning for UN staffing and other expenditure, rather than on value-creation).

Recommendation 9:

⇒ **DPPA should consider whether it can report a summary view of the cost/benefit provided by the conflict prevention, peacemaking and peacebuilding work of SPMs managed by the Department. An aggregated one-stop view of SPM achievements in accessing, engaging, and influencing relevant actors would help strengthen DPPA's value-for-money claim, even if these results must first be carefully de-identified and aggregated to avoid jeopardising ongoing peacemaking efforts.**

Finding : It is not possible to obtain a one-stop global view of the combined resources of key entities within the peace and security pillar, because reporting is divided among multiple UN entities, mandates, funding streams and reporting obligations. However, this Mid-Term Review identified interest from key stakeholders in gaining this kind of summary view, in a one-stop format .

Recommendation 10:

⇒ **DPPA should consider whether support for the UN peace and security pillar might be strengthened if a holistic view could be provided of the combined resources of SPMs, DPPA, the joint DPPA-UNDP programme, and UN Peacebuilding Fund.**

---

1. Objectives and scope of this review

The UN Department of Political and Peacebuilding Affairs (DPPA) commissioned an independent Mid-Term Review to assess:

i) the progress made by DPPA in the first 17 months of its 2020-2022 Strategic Plan, from 1 January 2020 to 31 May 2021, and
ii) how well DPPA's strategic planning tools have served the Department to date.

The review was requested to include concrete recommendations for improvements to both implementation and results reporting, while also factoring in changes in operating context due to the COVID-19 pandemic.

This is the first review of its kind under DPPA's present Strategic Plan, which is itself the first DPPA Strategy statement since the Secretary-General introduced reforms to the peace and security pillar of the UN, merging the former Department of Political Affairs (DPA) and Peacebuilding Support Office (PBSO) into one new Department, and reinforcing the close connections between the former DPA and the former Department of Peacekeeping Operations (DPKO - now renamed the Department of Peace Operations, DPO).

The **supporting research and analysis** for this review was conducted in June and July of 2021, relying primarily on review of documents and publications supplied by DPPA, interviews with senior DPPA staff from all Divisions, targeted additional document requests, and consultations with key external stakeholders.

The **Terms of Reference** for this review call for a stocktaking of DPPA's key achievements and gaps at the midpoint of the Department's three-year Strategic Plan, factoring in changes in operating context because of the COVID-19 pandemic, and assessing how well DPPA's strategic planning tools served the Department in this dynamic context. The Terms of Reference also request an examination of how the Department captures and reports on results, along with recommendations for improvements wherever this might be feasible. The Terms of Reference are focussed on DPPA's Strategic Plan 2020-2022, and do not extend to the Department's performance under the 'Strategic Framework' that accompanies the UN General Assembly's allocation of funds to DPPA under the UN regular budget.

The **primary audience** for the report is the Under-Secretary-General for Political and Peacebuilding Affairs, Rosemary A. DiCarlo, along with senior directors and staff within DPPA. The findings and recommendations of the Mid-Term Review are intended to serve as a basis for decision-making by DPPA leadership, in order to improve DPPA's performance and impact as outlined in the Strategic Plan. This report has been written with a broader audience in mind, including stakeholders, donors, counterpart organisations and researchers.

2. Implementation of the 2020-2022 Strategic Plan to date

2.1 A changing operational context for DPPA

DPPA's Strategic Plan 2020-2022 has been implemented in a **dynamic environment characterised by external and internal changes**. The trends and shocks affecting DPPA's work during the first half of the strategy period included ongoing geopolitical tensions, worsening conflict trends, constraints on UN resourcing and staffing, and the outbreak of the COVID-19 pandemic. These external changes were accompanied by multiple internal adaptations for the Department as it implemented the Secretary-General's reforms of the peace and security pillar. These elements are discussed below.

Geopolitical and conflict trends

DPPA's 2020-2022 Strategic Plan covers a period of **increasing tension** in international affairs, with the multilateral system under pressure, and the Security Council's response to armed conflict at times hindered by differences and tensions between its members. The rise of populist and authoritarian political leaders around the globe coincided with a decreasing level of trust between governments and their peoples and restrictions on women's rights and civic space. At the international level, rising geopolitical tensions continue to challenge international cooperation as envisaged in the UN Charter, including the collective security system.

Armed conflict saw increasing attacks on humanitarian workers and civilians amongst other breaches of International Humanitarian Law, along with drone warfare, indiscriminate bombardment, and cyber-attacks against civilian targets. Displacement caused by conflict and by environmental and economic factors moved waves of people across national and regional boundaries, while non-state actors exploited disorder through indiscriminate terrorist attacks, creating greater instability.

All of these tensions were exacerbated by the unchecked spread of the **COVID-19 virus** and the global pandemic in 2020-2021. The pandemic multiplied risks and previously existing inequalities and risks for vulnerable groups and populations affected by conflict, notably for women and girls, children generally (especially unaccompanied minors), detainees, displaced people and refugees, minority groups, the elderly and people with disabilities.

Constraints on UN Resourcing

The COVID-19 pandemic global outbreak in 2020 also coincided with constraints on the UN's funding position due to **delays in regular budget contributions** by some Member States which prevented the Department from filling a number of critical vacant positions.

Despite these constraints on hiring, DPPA continued to advance its peacemaking work, thanks to Member States' voluntary contributions to DPPA's Multi-Year Appeal fund, which received $35.9 million of the $40 million requested in 2020. The MYA funding mechanism enabled DPPA to maintain a flexible and rapid risk-response, continuing its conflict prevention, peacemaking and peacebuilding work, despite cuts made under the regular budget.

Reform of the UN peace and security pillar

Another major feature of DPPA's operational context in 2020 was the Secretary-General's three major reform tracks covering the peace and security pillar, development system, and management paradigm of the Organisation.

The peace and security reform involved the restructuring of DPA, DPKO and PBSO, as well as related cultural changes. A single political-operational structure under Assistant Secretaries-General with regional responsibilities, with dual reporting lines to the Under Secretaries-General for Political and Peacebuilding Affairs and for Peace Operations, link DPPA and DPO. The restructuring also included a merger of DPA and PBSO where the latter is to act as a 'hinge' connecting the whole peace and security pillar with the rest of the UN system, especially the development agencies, while still retaining a direct reporting line regarding the Peacebuilding Fund from the Assistant-Secretary-General for Peacebuilding Support, to the Secretary-General.

While the merger of formal structures entered into force on 1 January 2019, the implementation of these changes is an ongoing process at the time of reporting.

COVID-19 pandemic

The COVID-19 pandemic moved quickly **from a local, to national, and then global threat** in 2019 and 2020, placing pressure on already fragile relationships between populations and their governments, and between regional and global political powers. COVID-19 threatened to accelerate this erosion of trust, and to obstruct efforts to prevent and resolve conflict.

The outbreak of COVID-19 also presented a serious threat to electoral processes globally, with many countries rescheduling or postponing elections scheduled for 2020 at the national and local level, including a number of countries in which the UN was already providing electoral support in response to identified risks (such as national elections in Armenia, Bolivia, Ethiopia and Malawi; local elections in Papua New Guinea-Bougainville, Paraguay and Solomon Islands).

The pandemic presented additional challenges for UN peacemaking around the world as the ability of envoys and mediators to meet the parties, convene talks and travel was severely curtailed.

2.2 DPPA's response to the COVID-19 pandemic

The COVID-19 pandemic required DPPA to **rapidly adapt its working methods**. The Department adjusted its annual planning and reporting cycle during 2020 in response, requiring Divisions to plan and report on a quarterly rather than annual basis to promote rapid adaptation and reallocation of resources. Using the Strategic Plan to guide their planning, DPPA divisions rose to this additional challenge, developing quarterly work plans, which enabled DPPA to closely monitor the impact of the pandemic on the implementation of activities and the attainment of strategic objectives.

DPPA quickly **re-prioritized and re-allocated MYA funds** to match the needs in 2020-2021, reducing planned travel and staff deployments, and moving projects to modes of engagement compatible with the restrictions imposed by the COVID-19 pandemic. As a

result, DPPA's annual budget under the MYA was trimmed in 2020 from $45million to $40million, and it is expected to remain at that level until the end of the current Strategic Plan, in 2022. Staff interviewed for the purposes of the Mid-Term Review noted that it was fortuitous that the COVID-19 operational adaptations coincided with an unrelated freeze on UN hiring, meaning that DPPA's capacity to engage with relevant actors and networks was not entirely crippled by the constraints in filling vacant staff positions.

Identifying the potential for COVID-19 to exacerbate conflict patterns, DPPA supported the **analysis and response** of over 30 Special Political Missions (SPMs) and 100+ UN Country Teams, assessing the impact of COVID-19 on conflict dynamics, and proposing actions to foster inclusive peacemaking and prevent violence in the context of the pandemic. DPPA's COVID-19 risk analysis in 2020 was captured in a new risk register document,[1] which included DPPA's recognition of the impact of COVID-19 on the most vulnerable members of conflict-affected populations, frequently women or children (especially girls). In response, DPPA continued to push for implementation of the Women, Peace and Security Agenda, including through gender responsive analysis, targeted efforts to support women's meaningful participation and to address conflict-related sexual violence, and dedicated funding despite the constraints imposed by COVID-19.

DPPA's COVID-19 response included **briefings** to the Security Council on the peace and security implications of the pandemic, weekly briefings to the Secretary-General's Executive Committee on the political impact of COVID-19, together with the development of scenarios on the impact of COVID-19 outbreaks on local and regional conflict dynamics, in collaboration with the UN inter-agency Field Support Group on COVID-19. As part of this effort, DPPA also produced a tracker on the impact of the COVID-19 pandemic on the mandate implementation in UN field missions, including daily updates and periodic analytical notes. Increasing engagement with international financial institutions including the World Bank allowed DPPA to help shape global responses to the significant socio-economic impacts of the pandemic.

DPPA re-configured its services to enable the smooth **functioning of Security Council** processes and those of its subsidiary organs, despite the restrictions on travel and in-person meetings. This logistics and communication challenge was efficiently addressed, including the provision of online simultaneous translation, providing the Security Council and subsidiary bodies with a digital format for meetings that had only ever been conducted in-person previously. In-person field assessments for Security Council representatives were also replaced with immersive virtual briefings, allowing the Security Council to view the impact of ongoing conflicts using digital tools.

The **Secretary-General's appeal for a global ceasefire** on 23 March 2020 was supported by DPPA efforts in the UN Headquarters and in fragile and conflict affected areas, and subsequently received endorsement from over 180 Member States, as well as regional organisations, religious leaders, and a broad range of international and local civil society organisations.[2] In support of this drive for a global ceasefire, DPPA monitored developments on the ground, including steps to stop fighting, initial gestures of support, and unilateral ceasefires announced by conflict parties. A collaborative project with six academic institutions and non-government organisations provided a transparent view of the global

[1] See the sub-section on of part 3.1 of this report, on 'Risk Management and the Strategic Plan', below
[2] DPPA MYA Annual Report 2020

response to the Secretary-General's global ceasefire call, tracking key developments as they took place. [3]

While the call for a global ceasefire received strong endorsement from a broad range of actors, in practice some of the underlying drivers of conflict remained dominant, and prevented the call for a global ceasefire reaching its full potential. As noted by Under-Secretary-General Rosemary A. DiCarlo during an interview, the support of conflict-sponsoring powers was required in order for conflict parties to step back from the use of force during the COVID-19 pandemic.

Building on pre-existing efforts, DPPA promoted the safe **use of digital technologies** for conflict prevention and resolution. For example, DPPA was able to provide early training to mediators on digital process design and facilitation as they moved their operations on-line, and developed guidance on the use of social media in meditation. As a result, DPPA and SPMs successfully designed and implemented new hybrid models of mediation, combining inclusive in-person and digital interactions. As the advantages of these new hybrid models and digital tools become apparent, it is expected that they will likely become part of the prevention toolbox, even after the pandemic subsides. However, the Department also developed a clear awareness of those aspects of its prevention engagements which depend on direct engagement with counterparts.

DPPA also successfully pivoted to offer many of its **training and learning opportunities online**. It also increased its capacity to operate in conflicts involving the malicious use of digital technologies, by training UN staff on the use of good offices and other peacemaking techniques in conflicts where cyber capabilities are extensively used by the conflict actors and other parties, in addition to courses on issues such as drafting, conflict analysis, political economy analysis, and data-analytics.

DPPA in collaboration with UNDP, OHCHR, UN Women, UNESCO, UNOPS and WHO developed an operational guide on **conducting elections under COVID-19 restrictions**,[4] which served as a practical guide for UN electoral advisers. DPPA also reconfigured its interventions during 2020 and 2021 to minimise staff travel and exposure while continuing to provide support to these and other electoral projects in the field. This adaptation allowed DPPA to support the Under-Secretary-General to fulfil her role as the UN focal point for electoral assistance matters, pursuant to the mandate given by the UN General Assembly.

[3] The tracking tool can be accessed at: pax.peaceagreements.org/static/ covid19ceasefires. See DPPA Annual Report MYA 2020 at p.15

[4] See WHO, '*Public Health Considerations for elections and related activities in the context of the COVID-19 pandemic*' 10 December 2020, available at https://www.who.int/publications/i/item/WHO-2019-nCoV-elections-2020-1

2.3 Implementation of the Strategic Plan's risk-response model

The ultimate goal of DPPA's Strategic Plan 2020-2022 is to prevent conflict and sustain peace, underpinned by a rapid and flexible response to conflict risk. In pursuit of this goal, DPPA often acts in a supportive role, to influence settings away from violence. This approach is encapsulated by a '**risk-reduction' model** in DPPA's Strategic Plan.

In 2020-2021, DPPA demonstrated that it is well-positioned to fulfil its conflict prevention and sustaining peace mandate, launching flexible and timely risk-responsive initiatives thanks largely to the voluntary contributions of Member States under the Multi-Year Appeal (MYA).

For example, in 2020 alone, **DPPA's Standby team of mediation advisors** deployed on over 95 occasions, in approximately two-dozen contexts, despite the constraints imposed by the COVID-19 pandemic.[5] The DPPA Mediation Standby Team can be deployed anywhere in the world within 72 hours, addressing a range of issues related to peace negotiations. This enables DPPA to effectively support other UN bodies, UN Country Teams, and regional or sub-regional organisations regarding peacemaking or conflict prevention needs.

Taking into account other engagements of DPPA staff and advisors, DPPA successfully **deployed teams or individual experts** (virtually and in-person) a total of 188 times during 2020 in response to requests, an increase of more than one third compared to 2019 (139 deployments of staff or advisors). On average, this means that DPPA was providing risk-responsive interventions for conflict prevention and peacemaking more than 15 times each month. And in 72 per cent of cases in which DPPA received a request for electoral assistance in 2020, DPPA was able to field a coordinated response within four weeks, in line with response times for electoral support in 2019, despite the COVID-19 pandemic.

The available evidence for the Mid-Term Review of the DPPA Strategic Plan 2020-2022 suggests that the Department's risk-responsive posture continues to deliver 'impact on the ground' as intended by its Strategic Plan. DPPA's strategic risk-response approach forms an essential part of the Department's contribution to peacemaking globally.

2.4 Implementation of DPPA's strategic objectives

DPPA is performing soundly in the implementation of its strategic goals, based on the evidence available for this Mid-Term Review of the 2020-2022 Strategy. In 2020, the Department reported that it met or exceeded more than 79 per cent of its own performance measures under the Strategic Plan, and is on track to deliver similar performance in 2021. [6]

This success rate would arguably be higher if some performance measures which fall outside DPPA's control were removed from the tally. Excluding those results which fall outside DPPA's sole or primary influence, this review estimates that the Department is meeting or exceeding its targets for 81 per cent of performance measures under **Goal 1** (on conflict prevention and sustaining peace); 87 per cent of performance measures for **Goal 2** (on

[5] DPPA Annual Report MYA 2020 at p.25
[6] See DPPA Annual Report MYA 2020 at p.9

partnerships) ; and 87 per cent of performance measures for **Goal 3** (on institutional effectiveness). Given the constraints imposed on the Department in 2020 and 2021 by the COVID-19 pandemic, this is a significant achievement.

Table: DPPA's performance against strategic goals:

Reported performance to date	**DPPA Ultimate Aim: Reduce the risk of violence, promote sustained peace**[7]	
DPPA records show 79 per cent of annual performance measures were met or exceeded in 2020	***Three Main Goals***	***Seven Sub-Objectives***
	1. Contribute to preventing and resolving violent conflict and building resilience	1.1 Action-oriented analysis 1.2 Inclusive peace-making 1.3 Catalysing sustained peace[8]
	2. Strengthen partnerships for conflict prevention and resilience	2.1 Support to UN bodies and organs 2.2 Strengthened partnerships at the regional, national and local level[9]
	3. Achieve a learning, innovative working culture that takes forward the vision of the Secretary-General	3.1 DPPA is a learning, innovative and flexible department 3.2 DPPA has a collaborative work culture and an enabling work environment

3. DPPA's Strategic Planning Tools in Practice

3.1 The Strategic Plan 2020-2022 in practice

Finding : Evidence reviewed for this Mid-Term Review indicates that the Department's 2020-2022 Strategic Plan is useful in aligning DPPA effort, clarifying strategic logic through the Department's *'Theory of Change'*, and communicating the value of DPPA's work to outside audiences.

Recommendation 1:

⇒ **DPPA should consider developing a more operationally-focussed '*Theory of Change*' for its next strategy cycle, better reflecting the process through which DPPA**

[7] Summarised from the DPPA 2020-2022 Strategic Plan.

[8] Author's paraphrase. The original language from the DPPA Strategic Plan 2020-2022 frames sub-objective 1.3 as 'Sustained Peace: DPPA's peacebuilding engagements across the pillar and UN system catalyse efforts to address socio-economic and other grievances and risks. They are undertaken in partnership with Governments and relevant actors such as the World Bank and other international financial institutions. Sustainability informs priority areas of support to dialogue and coexistence initiatives, peace processes, and basic services'. See DPPA Strategic Plan 2020-2022 at p.22-23.

[9] This process is described as 'Expanding and deepening its (DPPA's) engagement regional and sub-regional organisations, international financial institutions and other stakeholders, as well as with Resident Coordinators and UN Country Teams'. DPPA Strategic Plan 2020-2022 at p.24

identifies risks of conflict, reaches relevant actors and networks, engages them in dialogue, and exerts influence for peace. Additional benefit might also be obtained if each Division were to formulate a 'Theory of Change' or strategic logic as part of its workplan, supporting the whole-Department strategic logic.

DPPA's 2020-2022 Strategic Plan is the Department's first statement of strategy since the structural reforms introduced to the peace and security pillar by the Secretary-General came into force on 1 January 2019. If successful, the Strategic Plan will be seen to have communicated the mission, vision, and values of DPPA, setting clear objectives, establishing priorities, aligning effort and resources, and affirming the significance of the Department's peacemaking, conflict prevention, and peacebuilding work.

Evidence reviewed for this Mid-Term Review indicates that **the Department's 2020-2022 Strategic Plan is serving its purpose well**. Senior managers and staff within DPPA state that the document is useful in aligning effort and clarifying the strategic logic through the Department's 'Theory of Change'. Some staff use the document as a ready reference for linking new initiatives, division work plans, and projects with existing strategy commitments, while others refer to the document primarily at DPPA reporting milestones. All staff interviewed agreed that the document is useful for communicating the value of DPPA's work to outside audiences, and some also commented that the Strategic Plan has proven useful in building a sense of shared identity among staff.

While the DPPA Strategic Plan does of course serve a 'political' purpose in promoting Member State support for DPPA, it is fundamentally more important that the Strategic Plan helps to ensure the strategic alignment of DPPA's own Divisions, and of the senior managers who direct them. This Mid-Term Review identified a risk that some senior managers within DPPA may regard the Strategic Plan as primarily a document intended for outside audiences, rather than being an operationally-focused document to be used in navigating DPPA through complex environments towards its key objectives. This risk is discussed further in the section below dealing with reporting obligations under the Strategic Plan.

The strategic logic of DPPA (Theory of Change)

For those DPPA managers who see the Strategic Plan as primarily for external audiences, greater levels of engagement might be obtained by emphasising those elements of the Plan that relate to operational priorities and decision-making, rather than internal capacity-building: the 'impact on the ground' of which the Strategic Plan speaks. In its next strategy period, DPPA might choose to perhaps direct more attention towards the central 'risk-response' logic also apparent within the current Strategic Plan, without neglecting of course the essential capability and culture elements of an effective organisation.

This would mean that a future iteration of DPPA's Theory of Change might consider directing less attention to DPPA's own resources, analysis, collaboration, partnerships, culture, learning, and innovation, which are primarily inwardly-focussed. While acknowledging the importance of these pre-requisites for organisational success, a more operationally-focussed Theory of Change might better reflect the outward-facing process through which DPPA identifies risks of conflict, reaches relevant actors and networks, engages them in dialogue, and exerts influence for peace, as summarised in the table below:

Table: Moving DPPA focus from internal to external.

2020-2022 Theory of Change / Strategic Logic	**Possible re-focus to more fully reflect DPPA's operational risk response**
« If DPPA deploys the full range of its **resources** based on cross-cutting **analysis**, in **collaboration** with others within the UN system and in **partnerships** with regional, national, and local stakeholders, drawing on an internal **culture** shaped by a commitment to **learning** and **innovation**, it will contribute to the prevention and resolution of violent conflict and to sustainable peace. »	If DPPA maintains strong **capability** (analysis, systems, partnerships, and a collaborative culture of learning and innovation), **it will be able to identify risks of conflict, reach relevant actors and networks, engage them in dialogue, and exert influence** for the prevention and resolution of conflict and more sustainable peace.
Focus of the argument: **Inward**. **DPPA's capability is the focus**, and there is no reference to operational engagements.	**Focus of the argument**: **Outward** **DPPA's operational engagements are the focus.** DPPA's capability is a prerequisite for success.
Author's paraphrase of 2020-2022 Logic: **Resources + analysis + collaboration + partnerships + culture + learning + innovation = DPPA's contribution** to the prevention and resolution of violent conflict and to sustainable peace.	**Alternative formulation:** **Capability + risk response + networks + engagements + influence = DPPA's contribution** to the prevention and resolution of violent conflict and to sustainable peace.

Aligning the strategic logic of DPPA Divisions

DPPA staff interviewed for the Mid-Term Review noted that while the Department has articulated its own Strategy and a Theory of Change at the whole-organisation level in the 2020-2022 Strategic Plan, additional benefit might be obtained if each Division were to formulate a corresponding theory of change or strategic logic at the divisional level, perhaps as part of its workplan. This would allow each Division to articulate a strategy relevant to the specific geography or mandate in question, based on sound conflict analysis and reinforced through collegial peer review, while demonstrating alignment with DPPA's global strategy and Theory of Change.

This process is likely to increase the level of 'ownership' of the DPPA Strategy at the divisional level, but would of course only work effectively if each Division's strategic logic remains congruent with the central logic of DPPA. In practice, there is little risk of the exercise introducing any divergence between the DPPA Strategy and the approach taken by

its divisions because the articulation of strategic logic will make starkly evident any differences, forcing a rapid and incisive strategic alignment.

Aligning DPPA & DPO strategy

Finding : The Secretary-General's 2019 reform of the UN peace and security pillar means that DPPA and DPO are now aligned by a common vision statement, collaboration between Executive leadership, a shared set of objectives, a shared regional structure, and an emerging shared risk management approach, while retaining their distinct mandates and responsibilities.

The Secretary-General's reform of the peace and security pillar in January 2019 created a *de facto* merger between the regional divisions of former DPA and DPKO, forming a single political-operational structure with regional responsibilities, guided by a common vision statement. Some DPPA staff interviewed for this Mid-Term Review noted that despite this re-structure, by June 2021 there was no common Strategic Plan for the two Departments that make up the peace and security pillar. DPPA is guided by its 2020-2022 Strategic Plan, while DPO follows a set of objectives entitled 'Secretary-General's Initiative on Action for Peacekeeping'.[10] While some staff see this as a missed opportunity to ensure strategic alignment between DPPA and DPO, others insist that the two Departments should not be expected to have a joint strategy, because they rely on separate UN mandates, resources and funding streams.

This Mid-Term Review noted however, that the two Departments are aligned through a common vision statement which lists key priorities, through strong collaborative leadership at the Executive level, and through a shared set of objectives and reporting framework under the Secretary-General's reforms, described as the 'Reform Benefits Tracker'.[11]

In 2021 DPPA and DPO are also working on a shared organisational risk assessment, which is intended to generate a joint risk register and risk treatment plan.[12] Given that the formulation of sound strategy is closely linked to the consideration of both risk and long-term vision, the Mid-Term Review concludes that the two Departments are already working to ensure strategic alignment in pursuit of their global mandates for peace.

Risk management and the Strategic Plan 2020-2021

Finding : The bulk of attention in DPPA's May 2020 risk register is devoted to the COVID-19 pandemic, for good reason. In most cases the risk definition is very broad and incorporates multiple risk factors into a single risk, which is likely to make the assessment of gravity and likelihood more complex for DPPA staff and management.

[10] Secretary-General's Initiative on Action for Peacekeeping. See https://www.un.org/en/A4P/, and https://reform.un.org/sites/reform.un.org/files/vision_of_the_un_peace_and_security_pillar.pdf, at p. 1
[11] See https://reform.un.org/content/peace-and-security-reform , https://undocs.org/A/75/202, and https://reform.un.org/sites/reform.un.org/files/vision_of_the_un_peace_and_security_pillar.pdf
[12] DPPA Annual Report MYA 2020 at p. 63. The joint DPPA-DPO risk management initiative was not examined as part of this Mid-Term Review.

Recommendation 2:

⇒ **Noting the approval of the UN Secretariat-wide risk register in July 2020, and bearing in mind DPPA's ongoing work to assess and manage risks on an organization-wide basis and in conjunction with DPO,[13] additional benefit might be gained by DPPA systematically covering additional categories of organisational risk, and directing attention to the most significant identified risk factors.**

Acknowledging the **close link between strategy and risk**, especially given the disruption caused by the COVID-19 pandemic, this Mid-Term Review briefly examined DPPA's risk management approach.

The enterprise risk register provided by DPPA for this Mid-Term Review was dated May 2020, and relates to the portfolio of DPPA projects funded by the Multi-Year Appeal, although the principles would be relevant to the entire range of DPPA projects, including those funded through the regular budget. DPPA's enterprise risk management document for the MYA addresses the risks posed by the global outbreak of the COVID-19 pandemic, including COVID-related implications for DPPA's political, reputational, operational, managerial and financial position. These COVID-19 related risks were collectively ranked as 'critical' and 'expected' in May 2020, and a number of risk mitigation measures were set out.

The bulk of attention in DPPA's May 2020 risk register is devoted to the COVID-19 pandemic, for good reason. In addition to these COVID-related risks, the DPPA risk register acknowledges ongoing risks to DPPA's financial position, to the successful implementation of the Women Peace and Security Agenda (WPS), and to DPPA's three strategic goals. In most cases the risk definition is very broad and incorporates multiple risk factors into a single risk, which is likely to make the assessment of gravity and likelihood more complex for DPPA staff and management.

Less attention appears to have been directed towards the following categories of risks, which might be expected to feature more prominently in future enterprise risk registers of this kind:

- **Reputational risks** that might diminish DPPA's standing and effectiveness
- **Physical risks** to the security of staff, consultants and equipment
- **Operational risks** preventing the completion of planned activities
- **Cyber risks** associated with DPPA's data and communication systems
- **Legal risks** encountered by DPPA when working across multiple jurisdictions[14]

Noting the approval of the **UN Secretariat-wide risk register in July 2020**, and bearing in mind DPPA's ongoing work in 2021 to assess and manage risks on an organisation-wide basis and in conjunction with DPO,[15] additional benefit might be gained by DPPA systematically covering all categories of organisational risk, and directing attention to the most significant identified risk factors. Ensuring a stronger approach to assessing and managing risks would also help advance the UN's peace and security reform.

[13] DPPA MYA Update 2021 at p.58
[14] Political risks are treated as assumptions in the current MYA risk framework.
[15] DPPA MYA Update 2021 at p.58

3.2 The DPPA Results Framework in practice

Finding : DPPA's robust and functioning Results Framework supports the Department's requirement for transparency, accountability, and a strong value-for-money claim under the 2020-2022 Strategic Plan.

Recommendation 3:

⇒ **DPPA should consider providing a one-page 'dashboard' view of its performance against strategic goals, combining both qualitative and quantitative assessments, while avoiding the temptation to reduce all of the Department's work to mere numbers.**

Recommendation 4:

⇒ **DPPA should improve on the relevance of indicators where feasible, focusing attention on DPPA's impact on the ground.**

DPPA's Results Framework serves multiple purposes, tracking the implementation of the 2020 - 2022 Strategic Plan, enabling reporting against the 'benefits tracker' associated with the Secretary-General's reforms, and providing data to other parts of the UN peace and security pillar for the purposes of their own reporting.

In addition to the challenge of these multiple end-use demands, DPPA's Results Framework must continue to meet the expectations of Member States by creating monitoring systems to substantiate the value of long-term and often intangible conflict prevention results such as trust-building, dialogue and peacemaking, in an evolving institutional context, and at increasingly frequent intervals. Interviews conducted for this Mid-Term Review universally acknowledged the difficulty of identifying and obtaining suitable evidence, especially regarding DPPA's conflict prevention objectives. This task is not straightforward, and DPPA is to be commended for embracing the challenge and creating a robust and functioning results reporting system to track the implementation of the 2020-2022 Strategic Plan.

The Mid-Term Review found that **DPPA's Results Framework supports the Department's requirement for transparency, accountability, and a strong value-for-money claim**. Without the Result Framework, DPPA would be solely reliant on anecdotal evidence drawn from specific interventions and peace processes, which are useful as illustrations and case studies, but do not provide a good basis for determining performance on a whole-organisation scale. The metrics and indicators in the Results Framework strengthen DPPA's reporting, and position DPPA well to continue demonstrating the value of the Department's global mandate for peacemaking and conflict prevention.

The **Results Framework effectively mirrors the DPPA Theory of Change** as expressed in the 2020-2022 Strategic Plan, helping the Department to maintain a coherent and persuasive statement of its effectiveness in complex environments. If DPPA were to re-formulate its theory of change to place more emphasis on operational results as suggested earlier in this report, then the structure of the Results Framework would also need to adapt to reflect these changes.

Scope for a dashboard view of performance

The current Results Framework was introduced in 2020 to track the implementation of the 2020-2022 Strategic plan, with baselines drawn from DPPA's operational data in 2019. The first 18 months of the Strategic Plan in 2020-2021 has provided DPPA with an opportunity to test what can be feasibly measured, and to identify those metrics that are most meaningful in practice. DPPA may now wish to consider consolidating some indicators to reduce reporting burden and data overload for readers, while maintaining the utility of the Results Framework for strategic purposes.

Where it is not possible to reduce the number of data points collected, DPPA's reporting of that data might still be streamlined by aggregating related measures together into a **simplified 'traffic light' or dashboard format for reporting**, while avoiding a simplistic over-reliance on quantitative indicators. This kind of aggregation would enable DPPA to provide a one-page view of its performance against each of its three strategic goals, focussed at the systemic level. Where an individual performance measure remains particularly salient for DPPA's strategy, the Department could of course direct greater attention towards that measure, as might be the case for data that demonstrates the risk-responsiveness that is foundational to DPPA's entire strategy.

To take the example of the Results Framework's treatment of Goal 1 of the Strategic Plan, in place of the four-page view provided by the 20 performance measures in DPPA's Results Framework, DPPA might consolidate these metrics into an aggregated and simplified view with only five traffic light indicators to help provide a dashboard view of areas where there may be an emerging risk of under-performing:

Text Box: Example of a dashboard report format for Goal 1 of the Strategic Plan

Goal 1: Prevent and resolve violent conflict and build resilience

DPPA's conflict prevention peacemaking and peacebuilding work is:

1. Informed (% of relevant targets attained or on track)
2. Inclusive (% of relevant targets attained or on track)
3. Responsive (% of relevant targets attained or on track)
4. Sustainable (% of relevant targets attained or on track)

By providing a snapshot view of this kind for all three goals, DPPA may be able to generate a one-page summary of performance, and reduce the burden of communicating the Department's performance under the Strategic Plan. This may also help build support to move DPPA towards a forward-looking adaptive system that confirms at intervals whether or not the Department is still moving in the right direction, rather than documenting in detail the Department's recent history.

Improving on the relevance of indicators

DPPA's Results Framework contains 48 performance measures that draw on 57 separate indicators (or 69 indicators if one counts the various sub-categories of gender-related metrics). Of these 57 indicators, 36 are raw numbers showing the total number of 'outputs' produced in a given year, such as the number of seminars, workshops, publications, visits, or contacts with relevant stakeholders. **DPPA complements these quantitative measures with extensive qualitative reporting**, including case study examples, to ensure that readers understand the subtlety of DPPA's discreet peacemaking, peacebuilding and conflict prevention role.

While continuing to use these quantitative indicators, DPPA should exercise caution to ensure that incentives are not created for inefficiency by treating increased activity as a measure of value in itself. As DPPA progresses further in its strategy implementation, some of these raw numbers could be improved upon by relating them more strongly to DPPA's strategic logic, using relative measures such as percentages, degree of coverage, or change over time, especially regarding risk responsiveness, networks, timeliness, quality, or flexibility.[16] Where certain 'raw' numbers have proven useful in communicating the work of DPPA to stakeholders in the past, these indicators should be maintained, even if they are imperfect.

More priority or attention could be directed to DPPA's impact on the ground, and less towards internal UN processes or the operational context itself. Some indicators featuring in the DPPA Results Framework focus attention on the context, or on factors outside the Department's control, rather than on the value created through DPPA's work. For example, indicators that assess the gender composition of conflict-party delegations in mediation processes are not in DPPA's power to control. DPPA has been working with UN Women over the last year to support consultations on the UN WPS monitoring framework, which will strengthen efforts to better measure progress toward the WPS agenda.

Some reported **indicators might benefit from being refined** to better measure the intended result. For example, DPPA reports on the level of satisfaction expressed by Member States benefitting from the services provided by the Security Council Affairs Division (SCAD), but the ultimate goal of the work carried out in this case is to ensure the procedural integrity and effectiveness of the Security Council Processes, which then helps reinforce the standing and authority of DPPA's peacemaking mission. Although one aspect of procedural integrity is indeed excellent service, the current 'client satisfaction' indicator could be complemented or replaced by a more direct and global measure of procedural integrity, such as an unqualified annual compliance audit of Security Council processes and systems against accepted criteria. If this approach were taken, SCAD would also be able to draw attention to the impressive progress made in ensuring timely publication of UN reporting on the practice and procedure of the Security Council in the *Repertoire.*

Some of DPPA's most valuable work goes beyond the three-year frame of the Strategic Plan. DPPA's role in conflict prevention requires the careful cultivation of relationships over the long term, often preceded by months or years of patient work to overcome reluctance

[16] Additional recommendations for refining indicators were advanced in the Value-for-Money Assessment of the DPPA MYA in November 2020. See https://dppa.un.org/sites/default/files/vfm_assessment_dppa_multi_year_appeal_1.pdf at p.36 and following.

on the part of key actors to engage with the UN, simply to gain access to the right people and networks. DPPA staff offered confidential examples of engagements, including one project in which four years of DPPA effort were required to obtain an open door with the right interlocutors, followed by two additional years of careful preparation by DPPA, which eventually allowed a senior UN representative to intervene based on signs of escalating conflict.

Significant outcomes such as documented peace agreements, unilateral declarations or ceasefires to which DPPA and the wider UN has contributed are such examples.

Highlighting hidden results

Finding : DPPA does not fully report on significant but hidden interim results such as the cultivation of trusted networks with key actors, and the quiet but substantial support provided by DPPA to UN Resident Coordinators, Country Teams, Development Coordination Offices, SPMs, and other partner organisations. Senior DPPA staff consider that these kinds of results are routine activities rather than results, and advised against seeking to quantify these aspects of DPPA's work.

Recommendation 5:

⇒ **Given the significance of interim results such as trusted access and engagement with the right actors, DPPA should examine whether it might be possible to report on the value of these hidden results more adequately, in an aggregated and de-identified manner, without jeopardising peace operations.**

In addition to more prominent and documented results, DPPA's work for peace generates valuable intangible or hidden results, whether interim or final. These include timely contacts with State and Non-State actors, the establishment of resilient networks and collaborative partnerships, the cultivation of discreet channels of communication with the right actors, support to UN Resident Coordinators and UN Country Teams, and collaboration with the other pillars of the UN. In order to achieve its intended 'impact on the ground', DPPA must invest intensively in creating trusted relationships, partnerships, and networks ahead of time, so that it is well positioned when a crisis arises or a conflict risk is identified.

Simply counting the number of these contacts or partnerships made by DPPA is not illuminating, but there is merit in better valuing DPPA's ability to create networks and 'anticipatory relationships' that encompass actors necessary for effective dialogue towards conflict prevention or peace. Despite the complex fragmentation of armed groups and conflict actors, each arena of conflict contains a limited number of key conflict actors, intermediaries, and supporters, so it should be possible to assess in rough terms the relative level of DPPA's preparedness and the sufficiency of the Department's networks and partnerships in each context.

Other significant but sometimes hidden results occur within the UN system itself, such as DPPA's work to ensure the procedural integrity of UN Security Council and subsidiary organ processes, or developing compromise language for UN Secretariat working papers on

sensitive subjects such as decolonisation. DPPA's in-house work to foster organisational learning, innovation, strategic planning, and evaluation also create significant, if sometimes hidden value for the Department's mandate. Staff interviewed for this Mid-Term Review affirmed the central importance of these ongoing efforts to build a learning, innovative, flexible and collaborative culture, as outlined under DPPA's third strategic goal.

In the field and when working with agencies from across the UN system, DPPA's hidden interim results include the essential work of the Department's staff to advise UN agencies regarding political pitfalls and sensitivities on the ground, and even mediating where needed between the different development actors present in conflict-affected environments. This critically important field work is accompanied by other valued results when DPPA quietly supports and accompanies high-level field visits by the Security Council and UN senior management, "backstops" SPMs,[17] advises on country-specific political processes bilaterally or through inter-agency task forces, or develops DPPA guidance and technical support for peace mediation teams in the field.

In addition, **DPPA creates significant value when it responds to requests for quiet support from regional organisations** carrying out mediation mandates, without seeking to encroach upon or compete with their lead role. This includes the Department's efforts to strengthen preventive dialogue with regional partners, and initiatives to expand the range of regional organisations and partners with which DPPA successfully engages in support of peace.

Noting the 2020-2022 Strategic Plan's statement that 'DPPA's success will be measured by impact on the ground'[18], and taking into account DPPA's underpinning risk-reduction model from the Strategic Plan, **there appears to be additional scope for recognising and reporting the value of DPPA's otherwise hidden results.**

3.3 The DPPA Annual Workplans in practice

Finding : The annual DPPA work plan process is rarely used for adaptation mid-year in response to changing contexts, although it has potential to serve this purpose if treated as a 'living document' owned by the Divisions. There is additional scope for DPPA Divisions to use the workplan for adjusting priorities and planned activities during implementation, rather than seeing the workplan process primarily as a reporting tool for donors. Despite the admirably concise nature envisaged for the workplan reporting template, DPPA faces the ubiquitous risk that the focus of operational reporting drifts towards reciting generic activities, rather than specific valued results.

Recommendation 6:

⇒ **To counter the natural tendency towards activity reporting, DPPA might consider requiring divisions to introduce each section with a headline statement and two-line**

[17] 'Backstopping' is a metaphorical term that derives from the net or barrier (the backstop) behind the batter in a game of baseball (USA), or the person standing behind the batter in a game of rounders (UK). The function of a backstop is to prevent the ball leaving the ground if it is not cleanly hit. In simple terms 'backstopping' means 'practical support and assistance', and it appears to be used as a 'catch-all' phrase. See https://dictionary.cambridge.org/dictionary/english/backstop

[18] DPPA 2020-2022 Strategic Plan at p.16

summary, to draw attention to the most valuable result generated, or the most significant risks identified and managed, and why this work was significant.

DPPA uses an annual workplan template to enable Divisions to prioritise work and report on progress, complementing the quantitative indicators collected against each Strategic objective through its Results Framework. At the end of 2020, DPPA harmonised the production of 2021 MYA project proposals and the annual workplan process, helping to align and streamline these related tasks. For the first time, this allowed the Department to get a more detailed, granular estimate of MYA funds required for each expected accomplishment under the Results Framework.

The onset of the COVID-19 pandemic forced DPPA to quickly adapt its strategic planning tools, including the annual workplan process. In 2020, DPPA management ensured a flexible adjustment to the risks of COVID-19 through a shortening of the usual annual planning and reporting cycle, which moved to a quarterly basis.

The workplan reporting template provided for this Mid-Term Review has two parts. The first asks DPPA teams to report against two issues of importance for DPPA by providing a short (up to 300 words) statement on progress in the implementation of the 'Women, Peace and Security' agenda, and another (up to 250 words) on the identification of risks and corresponding mitigation measures. The second part of the workplan template asks Divisions to report (in 2,000 words or less in total) against each of the seven strategic objectives under DPPA's 2020-2022 Strategic Plan, which allows for around 250-300 words per objective.

An impressionistic view of the contents and emphasis of DPPA workplan reporting can be gained from the wordcloud graphic below, compiled from all the divisional reports submitted against DPPA workplans for 2020.

Graphic: Informal 'wordcloud' representation of DPPA workplan reporting for 2020


Meet
Peace
UN
Region
Support
Risk
Engage
Women
Secure
Develop
Politics
DPPA
Team
Gender
Report
Virtual
Member
Active
Analysis
Key
COVID
Mission
Organs
Plan
Result
New
Council
Process
PMD
Project
Electoral
Action
Build
Help
Conflict
Inclusive
Peacebuilding
Partner
Internal
United Nations
Mitigate
General
Strategic
Participate
Committee
Joint
Society
Coordinate
Response
Culture
Cooperate
Enhance
Orient
Objective
Related
Body
Strategy
Sustain
Environment
Area
Strengthen
Assist
Inform
Prevent
Tool
Platform
Innovative
Pandemic
Dialogue
Policy
Group
Capacity
Elected
Enable
Impact
Training
Local
Ability
Event
Person
Time
PBSO
Brief
Civil
Led
Iraq
Country
Achieve
Rights
Flexible
Learned
System
Measure
Consulted
Review
Staff
State
Held
Remote
Africa
Partnership
Challenge
Agenda
APD
Collaborate
Effective
Share
Implement
Priority
Facilitate
Special
Fund
Recommended
Between
Situation
Increase
Discuss
Desk
EU
PBF
Early
Secretary
Central
Prepare
Address
Need
Adapt
Data
Contribute
Official
Regular
Context
Vaush
Expert
Network
Officer
Liaise
Ensure
Mediated
Social Assess
Cross cutting
Several
DPR
UNSG
Syria
Sudan
Draft
Focus
Travel
Skills
Ongoing
Possible
Lead
Regard
Note
Integral
RCO
Number
Ensure
Liaise
Officer
Expert
Network
Local
Person
Time
Brief
Envoy
Consulted
Review
Held
Remote
Staff
State
Several
DPR
UNSG
Cross cutting
Social Assess
Mediated
Syria
Sudan
Draft
Focus
Travel
Skills
Ongoing
Possible
Lead
Regard
Note
Integral
RCO
Number
Ability
Event
Person
Time
PBSO

The annual work plan is rarely used for adaptation mid-year in response to changing contexts, although it has potential to serve this purpose if treated as a 'living document' owned by the Divisions. **There is additional scope for DPPA divisions to use the workplan for adjusting priorities and planned activities during implementation**, rather than seeing the workplan process primarily as a reporting tool for donors. This may require adapting the workplan template and process to more closely resemble the informal working methods used by divisions to set, track, and adjust priorities on a weekly and monthly basis.

Despite the admirably concise nature envisaged for the workplan reporting template, DPPA faces the ubiquitous **risk that the focus of operational reporting drifts towards reciting generic activities, rather than specific valued results**. To counter this, DPPA might consider requiring divisions to introduce each section with a headline statement and two-line summary, to draw attention to the most valuable result generated, or the most significant risks identified and managed, and why this work was significant.

DPPA could potentially highlight the value of its partnerships and networks by placing additional emphasis on this section of the workplan report. These results capture the important work of the Department in cultivating trusted relationships, partnerships, and networks over the long term, including with conflict parties and relevant local, regional and multilateral actors, including the sometimes overlooked women-led elements of civil society. Other significant but sometimes hidden results such as the backstopping of SPMs could also be highlighted in the annual workplan reporting.

A strategically-focussed editorial process following the first submission of draft reports may help to DPPA staff to better highlight their key results, and more clearly state their own contribution by applying 'plain language' principles of writing. This should ideally:

- **Eliminate indirect language** (e.g. passive voice)
- **Introduce a 'so what?' statement** to explain why a reported result is significant to DPPA's peacemaking mission and strategy, and
- **Make a clear claim of DPPA's contribution,** in which Divisions describe the significance of DPPA's role in broad terms.

3.4 Overview of DPPA resources and their allocation in practice

Finding : It is difficult to obtain a summary of the financial resources available to DPPA in its peacemaking, peacebuilding and conflict prevention mission. DPPA's Strategic Plan extra-budgetary resources are obtained by the Department under the MYA, while its regular budget allocation from the UN General Assembly is managed separately. An overview of the various regular budget and extra-budgetary funds, and the total amount available to the Department would be a useful addition to DPPA's reporting, if this is feasible.

Recommendation 7:

DPPA should consider whether it is possible to provide a summary overview of the annual financial resources available to DPPA in its peacemaking, peacebuilding and

conflict prevention mission, including regular budget and extra-budgetary funding, and noting those resources that fall outside DPPA's Strategic Plan.

Finding : DPPA's allocation of funds under the MYA is aligned with the Strategic Plan.

Recommendation 8:

⇒ **DPPA should focus financial reporting under the MYA on value creation and 'return on investment', rather than the rate of expenditure of allocated funds.**

Finding : DPPA's reporting obligations are multi-layered. In addition to its reporting under the Strategic plan 2020-2022, the Department's core budget is subject to reporting under the UN regular budget framework approved by the UN General Assembly, which is not directly connected to the objectives set out in the Strategic Plan 2020-2022. Each year around $700 million is allocated from the UN regular budget to Special Political Missions (SPMs) managed by DPPA, along with $11 million in extra-budgetary funds. DPPA supports, guides, and oversees SPMs, but it does not report directly on their resources and results in the Strategic Plan Results Framework or in DPPA's own Annual Reporting. This means that the value-for-money offered by the SPMs managed by DPPA is not contained in a single report. DPPA's public MYA reporting already provides relatively detailed examples of selected SPM results in narrative form, but does not feature a summary cost/benefit overview. The SPMs report separately to the UN General Assembly (albeit in a format primarily focussed on results based budgeting and planning for UN staffing and other expenditure, rather than on value-creation).

Recommendation 9:

⇒ **DPPA should consider whether it can report a summary view of the cost/benefit provided by the conflict prevention, peacemaking and peacebuilding work of SPMs managed by the Department. An aggregated one-stop view of SPM achievements in accessing, engaging, and influencing relevant actors would help strengthen DPPA's value-for-money claim, even if these results must first be carefully de-identified and aggregated to avoid jeopardising ongoing peacemaking efforts.**

Finding : It is not possible to obtain a one-stop global view of the combined resources of key entities within the peace and security pillar, because reporting is divided among multiple UN entities, mandates, funding streams and reporting obligations. However, this Mid-Term Review identified interest from key stakeholders in gaining this kind of summary view, in a one-stop format.

Recommendation 10:

⇒ **DPPA should consider whether support for the UN peace and security pillar might be strengthened if a holistic view could be provided of the combined resources of the SPMs, DPPA, the joint DPPA-UNDP programme, and UN Peacebuilding Fund.**

DPPA Resources

This Mid-Term Review found that **it is difficult to obtain a clear overview of all the resources, both regular and extra-budgetary funds, on which DPPA relies.** The challenge of integrating different accounting systems, reporting methods, and lines of responsibility applied to the funds on which DPPA relies complicate the picture so that DPPA reports in different ways regarding different funds, without being able to present a consolidated view of the whole.

The **key sources of DPPA funding and their relevance to the DPPA Strategic Plan** 2020-2022 may be summarised as follows:[19]

- **The approximate $700 million annual 'regular budget' for the work of SPMs:** DPPA is the lead Department overseeing the SPMs and is closely involved in all major budgetary decision for these missions, but the mandate for SPMs is considered to be technically separate from DPPA's own Strategic Plan, because (with the exception of $11 million in extra-budgetary funds) the SPMs are funded by the UN regular budget allocation approved by the UN General Assembly. When viewed through the lens of UNGA mandates and UN funding approvals, senior DPPA staff noted that DPPA's Strategic Plan 2020-2022 does not apply in any way to the funding obtained from the UN regular budget for SPMs.[20]
- **The $45 million annual 'regular budget' allocation to DPPA:** Senior DPPA staff strongly advised that it would be technically incorrect to suggest these funds should be applied in accordance with DPPA's Strategic Plan 2020-2022, as the regular budget allocation is structured around a separate 'Strategic Framework' approved by the UN General Assembly. This separate framework translates the mandates assigned to DPPA into a programme of work, and loosely reflects the DPPA organigramme rather than DPPA's Strategic Plan 2020-2022. DPPA's regular budget allocation is considered as the 'core' funding mechanism for DPPA, and must be applied according to the terms approved by the UN General Assembly.
- **DPPA's $40 million Multi-Year Appeal fund (voluntary contributions made to DPPA by Member States):** The application of these extra-budgetary funds is tracked against the Department's high-level strategic objectives set out in the Department's Strategic Plan 2020-2022.

The UN Peacebuilding Fund and the UNDP-DPPA Joint Programme is related to DPPA's conflict prevention and peacemaking role, but does not feature clearly in DPPA's reporting on resources and results under the Strategic Plan 2020-2022, and this Mid-Term Review was not requested to examine the role of these funds.

The complexity in the allocation and accounting methods for DPPA's regular budget, extra-budgetary funds, and SPM funds is also reflected in the way that DPPA communicates publicly to stakeholders about its resources. The Department's quarterly and annual reports, and its updates on the annual $40 million MYA fund do not report on the $45 million obtained by DPPA each year via the UN regular budget, nor the more than $700

[19] See DPPA 2020-22 Strategic Plan, at page 31

[20] Funding for SPMs includes around $700 million UN regular budget allocated by the General Assembly (2020), supplemented by $11 million of extra-budgetary funds. See RB Budget reporting regarding SPMs, UNGA A/75/6 (Sect 3) 20-05968, 23 April 2020, at page 69ff.

million of regular budget funding annually for the peacemaking work of SPMs, each of which is approved by the UN General Assembly. For the time being, **there is no single DPPA report that provides a consolidated picture of the resources on which the Department relies, or how they are allocated under the Strategic Plan 2020-2022.**

Senior staff within DPPA advised that it would be inappropriate for DPPA to present a consolidated 'whole Department' view of resources and strategy implementation, because **the DPPA Strategic Plan 2020-2022 has no administrative authority over the regular budget funds on which the Department relies**. Each of DPPA's funding streams is provided on different terms, set by different actors: the MYA falls under DPPA's own authority, while the regular budget funding to DPPA and to SPMs relies solely on the mandate granted to DPPA by the UN General Assembly pursuant to Article 17 of the UN Charter. DPPA managers interviewed for this Mid-Term Review presented a variety of different views about whether the Department could or should present an integrated view of all resources applied to DPPA's peacemaking, peacebuilding and conflict prevention work, in particular regarding the significant work and resources of the SPMs.

Leaving aside the technical and administrative distinctions, this review does not recommend that DPPA should attempt to report in detail on the application of all of these various funds against each strategic objective under the Strategic Plan 2020-2022, for purely practical reasons. This effort would impose significant additional costs on the Department, for minimal benefit. DPPA has attempted to carry out this kind of reporting in the past, but the exercise required DPPA's operational staff to manually prepare timesheet reports showing how their time was allocated on a percentage basis to multiple relevant objectives, which was burdensome and inefficient, and delivered no operational or strategic advantage for the Department's work.

Despite the technical and organisational challenges, **it is clear that DPPA will be better positioned to successfully execute its strategy if it can align all available resources in support of that strategy**, and will then also be able to more clearly communicate a global picture of its value. If administrative or procedural obstacles remain insurmountable, the Department may of course choose to expressly exclude reporting on certain funding streams which are covered by the UN budget planning and reporting processes.

DPPA Reporting on resource allocation

If DPPA is to maintain its focus on conflict prevention, peacemaking and peacebuilding impact on the ground, **the reporting burden imposed on the Department's operational teams should ideally be minimised** to the extent possible, and wherever feasible, data that is also useful for operational decision-making should be prioritised over data which is compliance-oriented. At present, data is gathered to support reporting under several lines of accountability under the regular budget, the DPPA Strategic Plan 2020-2022, the Multi-Year Appeal, and the Secretary-General's reform benefits:

- ### Reporting on the use of DPPA's regular budget funds

  DPPA contributes to a separate data-collection and reporting process for the UN regular budget accountability requirements, which is conducted annually, gathering performance data regarding the preceding year.

Under the UN regular budget reporting system, results are grouped into five categories and 25 subcategories that unfortunately do not match the logic of the DPPA's Strategic Plan. A cursory review of these regular budget reporting categories shows that this system requires DPPA to quantify deliverables that relate to the internal UN machinery (e.g. delivery of seminars, workshops and training events) while much of DPPA's substantive work creating 'impact on the ground' is considered as unquantifiable under the regular budget system, and therefore effectively unreportable, including core DPPA functions such as consultations, advice and advocacy; good offices, fact-finding, monitoring & investigation missions, humanitarian assistance missions.[21]

DPPA is obliged to report under the UN regular budget system's 25 categories of deliverables, in addition to the reporting under DPPA's Strategic Plan, which features 48 different performance measures comprising more than 57 indicators. The end result of the regular budget reporting system provides an inadequate view of the peacemaking, conflict prevention, and peacebuilding contribution of DPPA, as can be seen from the compliance-oriented regular budget report of DPPA's performance in the 'prevention, management and resolution' of conflicts for 2019.[22]

As with most reporting systems that focus on compliance and accountability rather than strategic objectives, the UN regular budget reporting system appears to favour 'status quo' results in which the budgeted activities, expenditure, and deliverables are exactly as expected. Based on the evidence available to the Mid-Term Review, incentives for changing resource allocations under the regular budget appear to be minimal and beyond the control of DPPA.

Apart from the 'reporting fatigue' created for DPPA management and operational teams by the mandatory regular budget results reporting system, there is a risk that some DPPA Divisions may regard the DPPA's own Strategic Plan reporting as superfluous or as primarily directed towards donors, rather than being a strategic steering and adaptation mechanism. In the view of some senior DPPA staff interviewed for this Mid-Term Review, there is no link between DPPA's Strategic Plan and the $45 million regular budget resources obtained by DPPA each year, which weakens the authoritative influence of the Strategic Plan.

- ### Reporting on Special Political Missions

  Each year around $700 million is allocated from the UN regular budget to Special Political Missions (SPMs) managed by DPPA, along with $11 million in extra-budgetary funds. **While the work of the Special Representatives and Special Envoys of the Secretary-General is overseen by DPPA, this strategic effort does not feature fully in DPPA's Results Framework or the various DPPA-MYA Annual Reports and Updates**, despite this work representing some of the highest value peacemaking effort supported by the Department. This means that the reader is unable to form a

[21] Source: DPPA internal document, Explanatory presentation on regular budget categories of deliverables; See also UNGA A/75/6 (Sect 3) 20-05968, 23 April 2020, available at https://undocs.org/A/75/6(Sect.3)
[22] See page 15 of the DPPA report for 2019 and budget for 2021: UNGA A/75/6 (Sect 3) 20-05968, 23 April 2020, available at https://undocs.org/A/75/6(Sect.3)

view of the significant value-for-money offered by the SPMs managed by DPPA. DPPA's public **MYA reporting provides relatively detailed examples of selected SPM results in narrative form, but does not feature a summary cost/benefit overview**. The SPMs also report separately to the UN General Assembly (albeit in a format primarily focussed on budgeting and planning for UN staffing and other expenditure, rather than on value-creation).[23]

When these SPMs are seeking to avert the escalation of conflict in crisis situations, DPPA staff and senior management help develop contingency plans and scenarios, and participate in multiple related meetings and consultations. Despite the near-absence of SPMs from the Department's Strategic Plan, and the high levels of delegated authority conferred on Special Representatives and Envoys by the Secretary-General, **DPPA remains the Lead Agency for the SPMs, and supports, guides, and oversees their work.**[24]

The significance of the SPMs to DPPA in both operational and budgetary terms **calls into question whether and how DPPA should report on the work and results of the Special Envoys and Representatives of the Secretary-General within the Department's Strategic Plan**. At present these efforts are regarded as not being readily reportable by DPPA, as they are funded by a combination of regular budget and voluntary extra-budgetary contributions from Member States, each of which has a separate reporting system.

- ### Reporting on the DPPA Strategic Plan and the Peace and Security Reform

DPPA has effectively streamlined its reporting under the 2020-2022 Strategic Plan, the MYA, along with the peace and security reform benefits tracker, so that the information collected from operational teams is used for multiple different reports. During the first half of the 2020-2021 Strategic plan, information has been collected from operational teams each quarter, to support reporting under DPPA's Results Framework. This information also features in DPPA's six-monthly updates against its Results Framework, and in DPPA's quarterly and annual MYA Report.

Part of DPPA's Results Framework is also streamlined to provide performance information to the 'Reform Benefits Tracker' associated with the Secretary-General's reform of the peace and security pillar, which is owned by the Executive Office of the Secretary-General. Of the 48 performance measures within the DPPA Results Framework, 17 also appear in the Secretary-General's peace and benefits tracker.

- ### Reporting on strategic alignment in the use of MYA funds

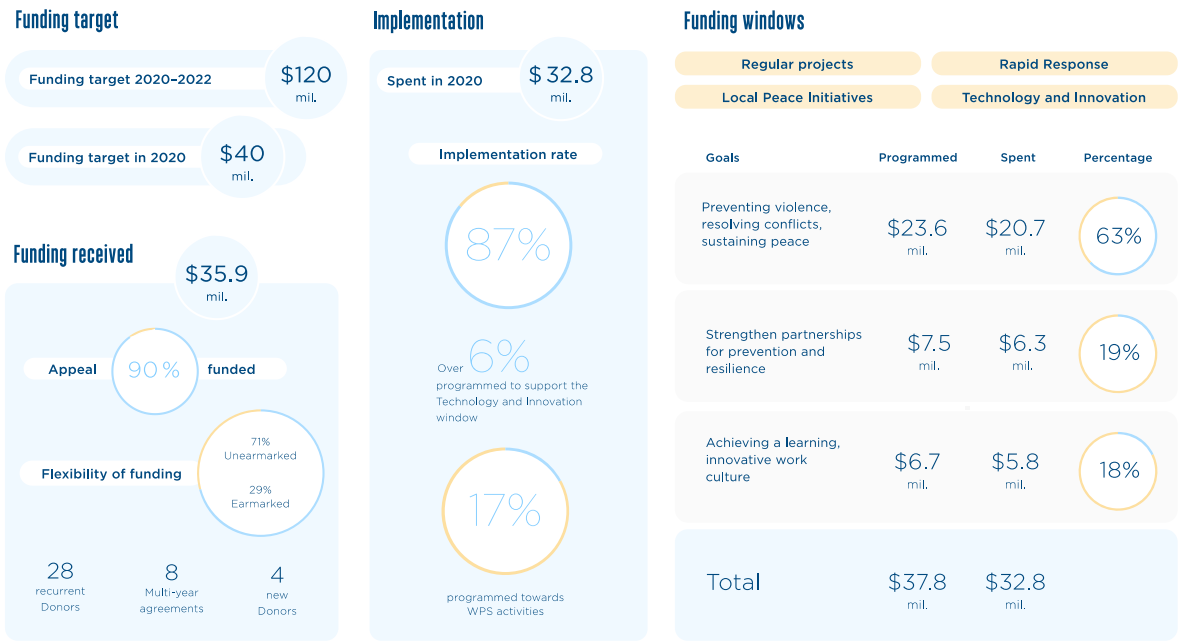
The alignment of MYA resources with the objectives of the Strategic Plan 2020-2022 is clear. DPPA reports on its performance against the objectives of the Multi-Year Appeal on a six-monthly and annual basis. When reporting on the use of MYA funds,

[23] See for example UNGA A/75/6 (Sect. 3)/Add.3

[24] See UNGA Doc A/75/6(Sect.3) Add.1 E at page 46 : 'Annex II: Lead department and mandates of special political missions 2021", and DPPA MYA Update 2021, at page 5.

DPPA uses the three goals in its Strategic Plan to show how resources are being allocated, as can be seen from the right-hand column of the graphic below.

Graphic: Reporting on strategic alignment of resources under the MYA[25]



However, DPPA's reporting on these MYA funds risks focussing attention on the percentage of funding allocated and spent (burn rate), rather than the results generated, as can be seen from the graphic above. Rather than reporting primarily on the 'burn-rate' of resources, DPPA might wish to focus more attention on the results attained.

In its annual report on the use of the extra-budgetary Multi-Year Appeal funds, DPPA also reports against total resources and expenditure over time, by year, and resources by donor, the assignment of Junior Professional Officers by each donor, and the split between MYA funding which is earmarked and unearmarked (i.e. constrained or unconstrained by a donor's conditions). The division between earmarked and unearmarked funding in the MYA helps show the extent to which DPPA is adequately resourced for an impartial, independent and timely response to risk, which is relevant to the objectives set out in the Strategic Plan 2020-2022.[26]

DPPA may wish to consider the potential for more precise reporting on the **cost/benefit of individual MYA projects**. This kind of project-level reporting might provide a view of how much is spent on each initiative within each division, within what timeframe, to achieve what result. When aggregated, this analysis would arguably allow DPPA to highlight outlying projects that are exceptionally cost-effective and nimble, while delivering valued results in a short time. At the other end of the scale, it may also help to identify initiatives which require extensive

[25] See DPPA Annual Report MYA 2020 at page 11.
[26] DPPA Annual Report MYA 2020 at page 69.

investment without yielding immediate results, but which may nevertheless have a strong value-justification over the longer term.

Given that DPPA manages a large portfolio of more than 100 MYA projects in each year of the 2020-2022 Strategy cycle, **a broad mapping of the entire portfolio might reveal additional patterns of cost/benefit and strategic alignment.** It may therefore be worth DPPA experimenting with prototype methods of mapping the portfolio of MYA projects against two or three simple criteria, such as cost, results achieved, or the speed of DPPA's response, even if the mapping remains approximate or indicative. This kind of method would require DPPA management to derive values for each project against its chosen criteria, which would be challenging, but not impossible if some degree of approximation is accepted.

While these portfolio mappings would remain indicative, they may still prove valuable in **helping DPPA senior management to highlight and communicate to stakeholders their own view of how DPPA's portfolio of projects is delivering significant impact for peace**. If DPPA were to begin reporting more on cost/benefit through portfolio mapping of this kind, this would arguably also help ensure strategic alignment and prioritisation at the whole-portfolio and Divisional level.

- ### Reporting on different financial instruments as a whole

Support for the UN peace and security pillar might arguably be strengthened if **a holistic view could be provided of the combined resources and activities of the SPMs, DPPA, the joint DPPA-UNDP programme, and UNPBF**.

UNDP and DPPA maintain a joint programme with its own fund. This 'Joint Programme' Fund has strategic objectives similar to DPPA's own, aiming to help Member States build national capacities for conflict prevention, but it is funded and reported on separately, and is not covered by DPPA's MYA fund or reports.[27] **The UN Peacebuilding Fund** also maintains its own reporting regarding its distinct fund, which again is different from DPPA's MYA fund. The **Department of Peace Operations** also maintains a separate extra-budgetary fund.[28]

While acknowledging the differences between the funds and programmes carried out by these related UN agencies,[29] this Mid-Term Review identified **some interest from key stakeholders in gaining a global view of how the UN works towards peace by using these different funding instruments**, how resources are applied and shared, and the combined results delivered by this collective effort.[30] There may be scope here to further advance the vision of 'One UN'.

[27] See the comparison between the various funds provided in *Multi-Year Appeal Update* document for 2019 at page 28, and the DPPA *Multi-Year Appeal for 2020-2022*, at p.42 ff
[28] See the overview provided in *DPPA Manual for the preparation of projects under the 2020 Multi-Year Appeal*, October 2019 at p.1. It fell outside the scope of this review to examine these funds
[29] For an explanation of the differences between the MYA fund, the UNPBF and the DPPA-UNDP Joint Program Fund, see the *Multi-Year Appeal Update* document for 2019 at page 28, and the DPPA *Multi-Year Appeal for 2020-2022*, at p.42 ff. Note that the PBSO extra-budgetary funding is not discussed in these tables.
[30] Source: Consultations carried out for this review. It was outside the scope of this Mid-term Review to consider this 'whole pillar' view.

4. Conclusion

At its midway point, DPPA's 2020-2022 Strategic Plan is demonstrably serving the Department well, and is being soundly implemented.

Despite a turbulent operational context, with increasing geopolitical tensions, troubling conflict trends, and the outbreak of the global COVID-19 pandemic, DPPA has shown its ability to rapidly pivot and adapt while continuing to implement its strategic objectives. The Department effectively adapted its programmes and processes to the reality of the COVID-19 pandemic and its constraints during the first half of the strategy period, re-allocating resources and applying new digital technologies to peacemaking, peacebuilding, and conflict prevention. DPPA has continued to apply the risk reduction model that lies at the heart of the Strategic Plan, and has delivered a strong performance against its own strategic objectives.

The second half of the 2020-2022 Strategic Plan offers DPPA the opportunity to further refine its strategic planning tools to remain focused on 'impact on the ground', to highlight valued achievements for peace including sometimes hidden interim results, and to ensure the alignment of resources and initiatives, both within the Department and with other agencies within the peace and security pillar. By continuing to strive for more effective peacemaking, peacebuilding and conflict prevention, DPPA upholds the vision of the United Nations as formulated by Member States in the drafting of the UN Charter: *To save succeeding generations from the scourge of war.*

Annex: Methodology

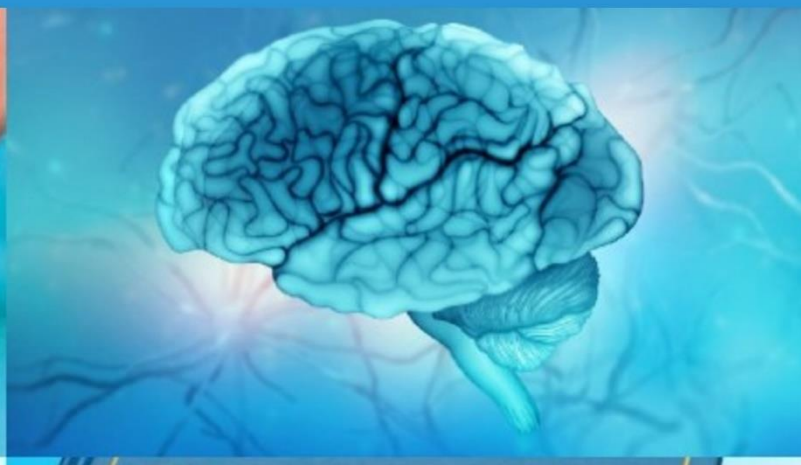
This Mid-Term Review was conducted in June and July of 2021 and designed using a light evaluation methodology involving document review, consultations, drafting, iterative feedback, and reporting. All information for the review was derived from existing documentary sources supplied by DPPA, and from interviews, as the scope of the review excluded analysis of original data. The review was guided by the principles of human rights and gender equality, and DPPA requested the reviewer to place special emphasis on DPPA's work related to Women, Peace and Security.

Consultations for this review included more than twenty interviews and meetings scheduled by DPPA, including with DPPA heads of divisions, the 45-member DPPA Planning Group, and DPPA senior leadership, along with multiple consultations with the DPPA Donor Relations team. Feedback received from key donors has also been incorporated in the report. The Mid-Term Review sought regular feedback from DPPA management during the conduct of the review and the reporting phase, to ensure the report remained focussed on issues of particular value for DPPA, including the functioning of DPPA's strategic steering tools in practice, and scope for improvements in those working methods.

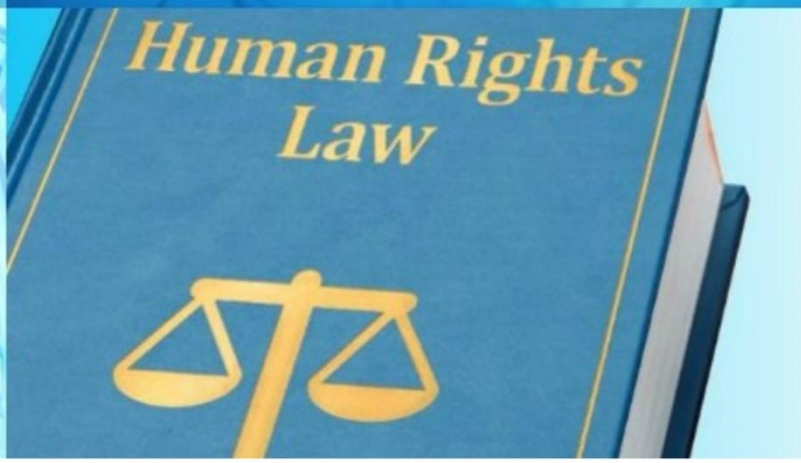
DPPA supplied relevant documents at the beginning of the review, which were then supplemented by additional documents and case examples provided by DPPA divisions.

Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025)









Adopted by the Committee on Bioethics (DH-BIO)
at its 16th meeting (19-21 November 2019)

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

COUNCIL OF EUROPE

STRATEGIC ACTION PLAN ON HUMAN RIGHTS AND TECHNOLOGIES IN BIOMEDICINE (2020-2025)

TABLE OF CONTENTS

INTRODUCTION 5

Vision and approach of the Strategic Action Plan 6

GOVERNANCE OF TECHNOLOGIES 8

Embedding human rights in the development of technologies which have an application in the field of biomedicine 8

Fostering public dialogue to promote democratic governance and transparency in the field of biomedicine. 9

EQUITY IN HEALTHCARE 11

Promoting equitable and timely access to appropriate innovative treatments and technologies in healthcare 11

Combating health disparities created by social and demographic changes in Council of Europe member States. 12

PHYSICAL AND MENTAL INTEGRITY 13

Strengthening children's participation in the decision-making process on matters regarding their health. 13

Safeguarding children's rights in relation to medical practices which have future or long-term implications for them. 14

Safeguarding the rights of persons with mental health difficulties. 14

CO-OPERATION AND COMMUNICATION 16

Developing long-term strategic co-operation with Council of Europe committees and other intergovernmental bodies working in the field of bioethics. 16

Ensuring the communication and dissemination of the outputs of the Committee on Bioethics to internal and external stakeholders in order to maximise their uptake and utility 17

IMPLEMENTATION 19

LIST OF SOURCES 21

INTRODUCTION

1. We are now at a turning point in human rights in biomedicine. This became evident during the International Conference that was held in Strasbourg on 24-25 October 2017 on the occasion of the 20th anniversary of the Convention on Human Rights and Biomedicine (Oviedo Convention), the only international legally binding instrument exclusively concerned with human rights in biomedicine. The Conference concluded that the principles enshrined in the Oviedo Convention remain of crucial relevance and that, in the 20 years since the Convention came into force, important new human rights challenges have emerged that need to be addressed.

2. Bioethics is often construed as a “culture of limits”. However, its role should be to accompany progress in science and to reflect on and to protect and promote human rights. Bioethics serves to safeguard human rights principles and goes to the heart of how we want to shape both the lives of individuals and the broader society. Human rights challenges are posed by scientific and technological developments as well as by the evolution of established practices in the biomedical field.

3. New technologies are emerging, for instance in the field of genetics, and some technologies, such as those involving artificial intelligence and big data, are being combined to produce new applications. The application of emerging and converging technologies in biomedicine results in a blurring of boundaries, between the physical and the biological sciences, between treatment and research, and between medical and non-medical purposes. Although they offer significant opportunities within and beyond the field of biomedicine, they also raise new ethical challenges related to inter alia identity, autonomy, privacy, and non-discrimination. The Committee on Bioethics has been discussing these emerging and converging technologies for some time and has developed considerable expertise in addressing the human rights challenges posed by them.

4. Important human rights challenges are also emerging through established practices in the field of biomedicine. Changes in the perception of the decision-making capacity in children, persons with mental health difficulties, and vulnerable older persons, are prompting reconsideration of the balance between protection and respect for autonomy. In addition, important demographic changes, such as migration and ageing populations, coupled with budgetary restrictions in healthcare, are resulting in new or increasing barriers to accessing healthcare services. At the same time, there is unprecedented scientific progress, which results in innovative therapies that are not always available or affordable to disadvantaged individuals and groups. This development indicates that, in addition to the traditional focus on patient’s rights, there is a need to guarantee equitable access to healthcare.

5. The Council of Europe is uniquely placed to address these developments through its Committee on Bioethics with regard to the Oviedo Convention, and has an important role in being a forum for continuous reflection and discussion to root the answers to new ethical challenges in human rights and shared European values.

Vision and approach of the Strategic Action Plan

The vision and approach of the Strategic Action Plan are to protect human dignity and the human rights and fundamental freedoms of the individual with regard to the application of biology and medicine. The Strategic Action Plan puts particular emphasis on addressing the challenges posed by new technological developments and by the evolution of established practices in the field of biomedicine.

6. In January 2018, a drafting group was established to elaborate the Strategic Action Plan. A number of drafting group meetings were held, the fruits of which were presented and discussed during plenary meetings of the Committee on Bioethics, notably in June and November 2018 and in June 2019. The feedback received from member States' delegations was incorporated in the Strategic Action Plan. To ensure synergy with others, information was provided to, and exchanges held with, a number of Council of Europe committees. There were also exchanges with a number of intergovernmental bodies as a means of developing long-term strategic co-operation. The Strategic Action Plan was adopted by the Committee on Bioethics at its 16th meeting in Strasbourg on 19-21 November 2019.

7. The Strategic Action Plan was developed by the Committee on Bioethics based on a number of preparatory studies, replies to questionnaires, and the findings of international conferences. The Strategic Action Plan also considers the work that has been done or that is currently under way in other Council of Europe committees and other intergovernmental organisations.

8. The Strategic Action Plan is built on four thematic pillars. Three of these pillars correspond to three critical human rights aspects that are affected by the new developments: governance of technologies; equity in healthcare; and physical and mental integrity. The fourth pillar is transversal and concerns co-operation and communication. These pillars contain strategic objectives and actions.

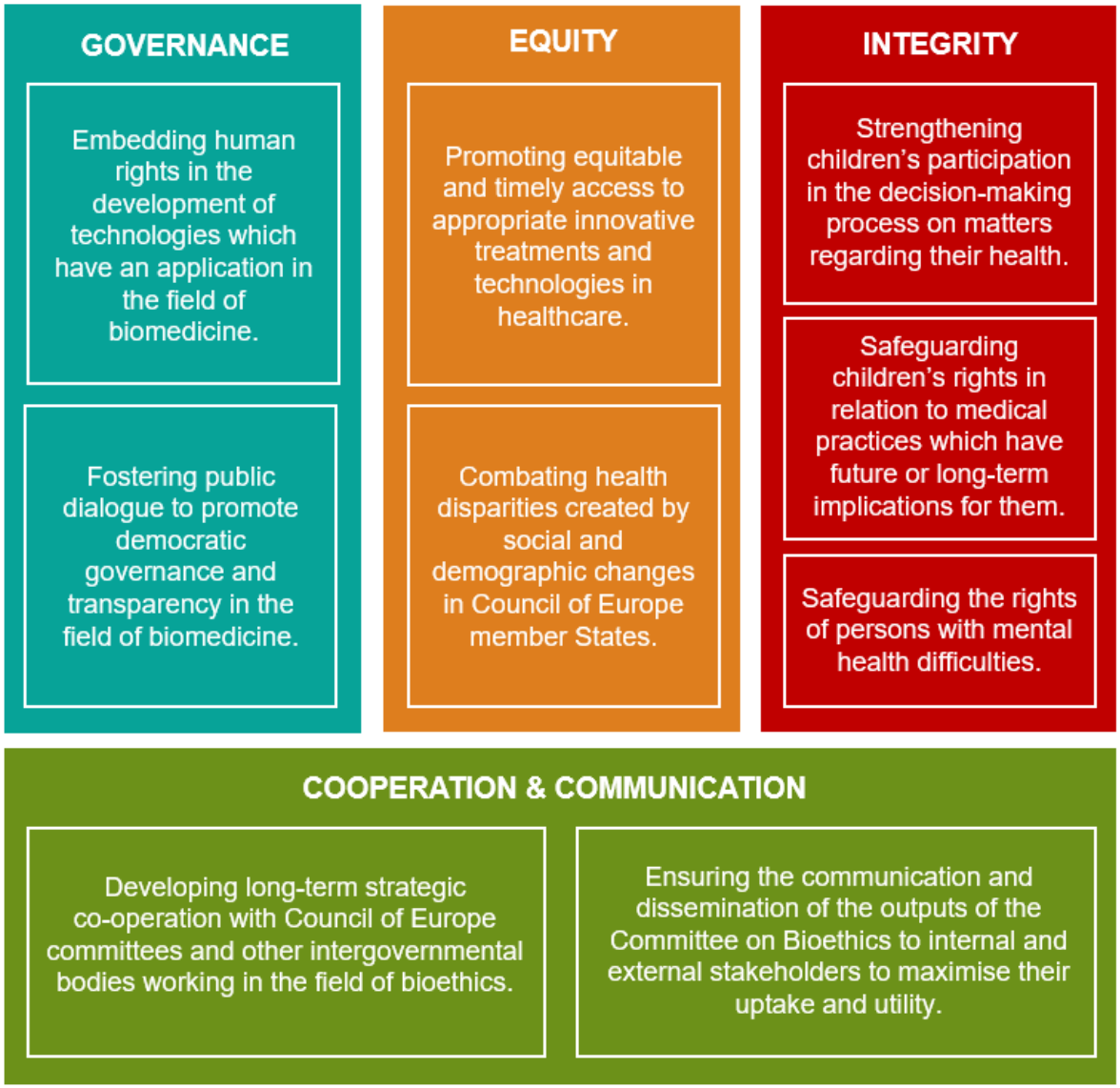
9. Priority actions for the 2020-2025 reference period were determined on the basis of several criteria, including the demonstrated need; the feasibility in light of available resources, expertise, and time; the impact on Council of Europe member States and their populations; the potential to elicit changes in policy or practice over the longer term; and opportunities to pool resources and increase impact through co-operation with the other committees of the Council of Europe and/or with other intergovernmental organisations. The range of activities has also been balanced to ensure that due attention is given to building on previous work by the Committee on Bioethics and to the implementation of previously elaborated tools.

10. The proposed actions take into account complementarity and co-operation with internal and external key partners. Several issues identified as posing important human rights challenges in the field of biomedicine, such as migrant health, have not been included because they are already being comprehensively addressed by other bodies. Further, it should be noted that various actions should be considered as building blocks for future work to be extended beyond the lifetime of the current Plan.

11. The timeline introduced at the end of the Strategic Action Plan outlines the expected year of delivery of the outcomes of the actions. In realising a specific action, the Committee on Bioethics intends to initiate work well in advance of the expected

year of delivery and a number of modalities will be determined, including the establishment of drafting groups, the commissioning of expert reports, and the organisation of seminars.

Structure and specific objectives of the Strategic Action Plan



GOVERNANCE OF TECHNOLOGIES

12. Research and innovation are particularly difficult to govern because they create novelty and surprise. Rolling out technology into society is a complex and unpredictable process. The full extent of the risks and unintended consequences of a given innovation can only be fully appreciated with experience, and by that time, control and change can be difficult, if not impossible, as the technology becomes embedded in social infrastructures or human culture. The ways in which technology is steered and controlled have significantly changed. Whereas before technology was governed mainly by national governments that adopted regulations to protect the rights and freedoms of citizens, new technologies are now governed in more heterogeneous and flexible ways by a variety of stakeholders, arguably with less focus on the protection of human rights.

13. Governance frameworks are necessary to optimise the chances of stimulating innovation that contributes to human flourishing, whilst minimising applications that have negative consequences for individuals and society. Therefore, the first pillar of the Strategic Action Plan addresses the governance of technologies, emphasising that it is necessary to change the way in which technologies with an application in biomedicine are governed. Governance models are required to guarantee that the protection of human rights is a guiding consideration throughout the entire process of research, development, and application. In addition, ongoing dialogue between the public, scientists, and policy makers should be ensured so that technological developments are robustly deliberated, democratic, and legitimate.

Embedding human rights in the development of technologies which have an application in the field of biomedicine.

14. Technological innovation often creates its own dynamic. Major technological breakthroughs in fields such as artificial intelligence, genome editing, and neurotechnology have the potential to advance biomedicine and healthcare. However, uncertainty exists about the impact and direction of these developments. For example, artificial intelligence is increasingly proficient in diagnostics but depends on massive amounts of patient data which may impact on transparency and patient trust, thereby necessitating the provision of guidance for healthcare professionals. Genome editing techniques which introduce inheritable changes in the human genome raise serious concerns about the possibilities of irreversible harm to future persons. Developments in neurotechnologies, such as deep brain stimulation, brain-computer interfaces, and artificial neural networks, raise the prospect of increased understanding, monitoring, but also of control of the human brain, raising issues of privacy, personhood, and discrimination.

15. The role of governance in biomedicine is often restricted to facilitating the applications of technology and to containing the risks that come to light. In this way, human rights considerations will only come into play at the end of the process, when the technological applications are already established, and the technological pathways often have become irreversible. To overcome this problem, there is a pressing need to embed human rights in technologies which have an application in the field of biomedicine. This implies that technological developments are from the outset oriented

towards protecting human rights. For that reason, governance arrangements need to be considered, which seek to steer the innovation process in a way which connects innovation and technologies with social goals and values.

Actions

- **Examining Article 13 of the Oviedo Convention in the light of developments in gene editing technologies.**

  In its statement of December 2015 on gene editing technologies, the Committee on Bioethics made a commitment to examining the ethical and legal challenges raised by genome editing technologies in the light of the principles laid down in the Oviedo Convention. To this end, this action necessitates an examination of the practical and legal implications of Article 13 of the Oviedo Convention as it relates to the use of gene editing technologies in the context of research, and of clinical applications of gene editing in somatic cells and the germline. The examination may indicate a need to clarify or amend Article 13.

- **Assessing the relevance and sufficiency of the existing human rights framework to address the issues raised by the applications of neurotechnologies.**

  Applications in the field of neurotechnology raise issues of privacy, personhood, and discrimination. It therefore needs to be assessed whether these issues can be sufficiently addressed by the existing human rights framework or whether new human rights pertaining to cognitive liberty, mental privacy, and mental integrity and psychological continuity, need to be entertained in order to govern neurotechnologies. Alternatively, other flexible forms of good governance may be better suited for regulating neurotechnologies.

- **Developing a report on the application of AI in healthcare, in particular regarding its impact on the doctor-patient relationship.**

  Artificial Intelligence (AI) has the potential to improve diagnostic and therapeutic outcomes for patients. Although deep learning algorithms in a variety of tasks in radiology and in medicine generally have demonstrated significant promise, it is likely to be several years before AI is mainstreamed into the healthcare domain. The predictive capability of AI raises concerns about privacy and discrimination. Moreover, as AI evolves, it will create new complexities for the doctor-patient relationship. In the light of these challenges, the Committee on Bioethics intends to prepare a report highlighting the role of healthcare professionals in respecting the autonomy, and right to information, of the patient, and in maintaining transparency and patient trust as critical components of the therapeutic relationship.

Fostering public dialogue to promote democratic governance and transparency in the field of biomedicine.

16. In order to guarantee that the directions of innovation and the ethical challenges raised by technological developments are robustly deliberated, governance should go hand in hand with broad and informed public dialogue. Fostering a dialogue between the public, scientists, and policy makers should promote democratic governance and

transparency in the field of biomedicine. This can assist policy makers in public consultations and, therefore, in ascertaining the most appropriate governance models needed for biomedical technologies and their applications. This is in line with Article 28 of the Oviedo Convention, which provides that "the fundamental questions raised by the developments of biology and medicine are the subject of appropriate public discussion in the light, in particular, of relevant medical, social, economic, ethical and legal implications, and that their possible application is made the subject of appropriate consultation".

Actions

- **Translating the Guide to public debate on human rights and biomedicine in non-official languages and disseminating it in Council of Europe member States.**

  The Guide to public debate, presented at the high-level seminar held in Strasbourg on 4 June 2019, is a tool for policy makers to help them engage with the public. It aims at raising public awareness, promoting discussion between different actors, groups, and individuals, including those who are marginalised and disadvantaged, and at facilitating consultation of the public by authorities with a view to making policy decisions. Translating the Guide to public debate into non-official languages and disseminating it will foster public debate initiatives in Council of Europe member States, including in countries and regions where public debate is less developed.

- **Promoting dialogue amongst the public, practitioners, and policy makers to ensure that patient and public interest is a key priority in the development and regulation of genomic medicine.**

  The future success of personalised medicine depends upon access to and sharing of exceptionally large amounts of genomic and other health data from patients and healthy individuals. The concept of solidarity recognises our common vulnerability to illness, and that we will all need healthcare at some point in our lives. Solidarity emphasises the willingness to accept certain potential costs (e.g. sharing our genetic data) in order to realise the common good, in this case better healthcare. Altruism and solidarity are intertwined with the principle of reciprocity. In agreeing to share genetic information, this gives rise to certain obligations on the part of researchers, healthcare professionals, and the state. These include providing information to data donors, including in relation to incidental findings, robust governance mechanisms, and equitable access to the treatments developed. In the interests of patients and the general public, the Committee on Bioethics intends to promote a dialogue between the public, practitioners, and policy makers on how to incorporate the principle of reciprocity in the governance of genomic medicine.

EQUITY IN HEALTHCARE

17. Since the adoption of the Oviedo Convention, developments in biomedicine and in society have taken place that result in increasing disparities in access to healthcare. For instance, an increasing number of innovative treatments and healthcare technologies have entered the market yet, because of their price, may not be accessible to everyone. In a parallel development, broader social and demographic changes (e.g. ageing populations and migration) are causing some groups in society to systematically face more difficulties in accessing healthcare. These difficulties are compounded by budget cuts which are putting pressure on healthcare systems and are increasing the risk of inequities in healthcare. These inequities are especially harmful for individuals and groups who are already disadvantaged.

18. The second pillar of the Strategic Action Plan addresses the increasing risk of health disparities by promoting equity, in accordance with the right to equitable access to healthcare pursuant to Article 3 of the Oviedo Convention. This obliges States party to the Convention, to adopt the necessary measures to prevent discrimination, thereby implying the identification, reduction, and ultimately elimination of disparities in access to existing and new treatments and technologies. This necessitates special efforts to improve access for disadvantaged individuals and groups, and to ensure that new developments do not create or exacerbate existing disadvantage.

Promoting equitable and timely access to appropriate innovative treatments and technologies in healthcare.

19. New developments in healthcare hold the promise of greatly improved health but can entail, at the same time, risks of deepening inequalities and new forms of discrimination and marginalisation. For instance, innovative treatments, such as for cancer, multiple sclerosis or very rare medical conditions, are often expensive and may only be affordable to a small portion of the population. Similarly, new healthcare technologies, such as health apps, telemedicine, and healthcare assistive robots, may only be available to those who possess the knowledge, skills, and financial means to use them. Consequently, it is necessary to encourage member States to ensure that new treatments and healthcare technologies are made available in an equitable and timely manner.

Action

- **Elaborating a draft Recommendation on equitable and timely access to innovative treatments and technologies in healthcare systems.**

  It is essential that innovative treatments and new healthcare technologies are made available in an equitable and timely manner. However, in view of the competing demands on healthcare services, it may be a challenge to know how best to achieve this goal. To assist member States, the Committee on Bioethics intends to prepare a Recommendation laying down principles to ensure that patients may benefit from timely and affordable access to safe and effective medicines, and that fairness and consistency in decision-making, regarding equitable access to the products of innovation, are promoted.

The Recommendation, while allowing flexibility at member State level, would ensure that decisions regarding access to innovative treatments and interventions would take account of fundamental principles such as justice and beneficence. Moreover, a harmonised framework across member States would help to combat inequities between them and to empower them. This is especially relevant considering that many citizens travel between states to access innovative treatments and technologies, which is a challenge for all member States.

Combating health disparities created by social and demographic changes in Council of Europe member states.

20. There is concern that existing healthcare resources are less accessible to certain patient populations because of their particular social circumstances which can make it a challenge for them to access, for example, valid health information and appropriate care. More specifically, the issue of equitable access to healthcare for persons in vulnerable situations is an enduring challenge for member states. For instance, access to clinical trials and innovative treatments and healthcare technologies often depends on information found on the internet and social media which may be more difficult for them to glean. Combating such health disparities is therefore important, for instance by making healthcare services and resources more accessible and by training healthcare professionals to ascertain their level of health literacy and capacity to participate in decision-making.

Action

- **Developing a Guide to health literacy for persons in vulnerable situations in order to empower them to access health care of appropriate quality on an equitable basis with other groups in society.**

  It has been well documented that older persons experience difficulties in exercising their right to access health care services but this is a concern which applies to a broader scope of "persons in vulnerable situations" and therefore potentially anybody not necessarily belonging to an identified group nor being a patient.
  This has become even more challenging as a result of the emergence of innovative treatments and new healthcare technologies that are very expensive and may require specific knowledge and skills to obtain. At the same time, established practices in healthcare have become more patient-centred and attentive to human rights, in a way that increasingly recognises the rights and decision-making capacity of persons in vulnerable situations. To this end, it is essential that they understand health information and know what healthcare services are available and how best to access them. In response to this need, the Committee on Bioethics intends to prepare a Guide to health literacy for equitable access to healthcare in order to empower them to be more effective advocates in accessing healthcare services and in making appropriate decisions regarding their health.

PHYSICAL AND MENTAL INTEGRITY

21. Technological developments in the field of biomedicine create new possibilities for intervention in individual behaviour. For instance, certain technologies raise the prospect of increased understanding, monitoring, and control of the human brain, while other developments allow for the permanent health monitoring of individuals. These developments raise novel questions relating to autonomy, privacy, and even freedom of thought. Moreover, the evolution of existing practices, such as the collection and sharing of genomic and health data, may give rise to similar concerns. There should also be consideration of other important social trends (e.g. pressure of social media on young people) and changing societal perceptions in how to balance the protection and respect for autonomy of children, persons with mental health difficulties, and vulnerable older persons, with increased recognition of their decision-making capacities.

22. In the light of these developments, the third pillar of the Strategic Action Plan addresses concerns for physical and mental integrity. Guaranteeing respect for a person's integrity in the sphere of biomedicine is one of the central tenets of the Oviedo Convention. This is understood as the ability of individuals to exercise control over what happens to them with regard to, inter alia, their body, their mental state, and the related personal data.

Strengthening children's participation in the decision-making process on matters regarding their health.

23. There are changes in the general perception of the autonomy and protection of children regarding their capacity to participate in decision-making. This is confirmed and endorsed by human rights instruments, notably the UN Convention on the Rights of the Child, which recognises that children are rights holders with a progressively evolving ability to make their own decisions. However, on matters concerning their health and general well-being, there is uncertainty as to how the increased recognition of their decision-making capacity should be addressed. Finding the right balance between autonomy and protection is a challenge when considering that children's rights are situated within a larger set of parental rights and responsibilities which also focus on their best interests.

Action

- **Developing a Guide to good practice concerning the participation of children in the decision-making process on matters regarding their health.**

  Acknowledging the need to recognise the evolving nature of the decision-making capacity of children also in matters regarding their own health, the Committee on Bioethics intends to prepare a Guide, containing principles and good practices, to involving children in medical decision making. This will include consideration of the rights of the child, the rights and responsibilities of the child's legal representatives, and the child's interests interconnected with those of their family members. The Guide should primarily target healthcare professionals but should also be accessible to the children's parents and/or legal representatives.

Safeguarding children's rights in relation to medical practices which have future or long-term implications for them.

24. Every child is a rights holder in his or her own capacity as recognised in Article 14 of the UN Convention on the Rights of the Child. The child's autonomy can be conceptualised as "the child's right to an open future", meaning a right to have one's future options kept open until one can make one's own decisions. The content of the right to an open future therefore includes restrictions on what parents (and others) can do for children, and, on some interpretations, indicates what parents (and others) ought to provide children with. There are challenges regarding the most appropriate interventions which parents and others should be allowed to authorise in order to safeguard the health of the child.

Action

- **Organising a seminar on relevant legislation and good practices with regard to early intervention on intersex children.**

  Resolution 2191(2017) of the Parliamentary Assembly of the Council of Europe on promoting the human rights of, and eliminating discrimination against, intersex people, calls for "medically unnecessary, sex-"normalising" surgery" on intersex babies to be prohibited, along with other treatments practiced on intersex children and young people without their informed consent. It recommended to carry out further research into the long-term impact of these treatments and to ensure that, unless there is an immediate risk to the life of a child, altering the sex characteristics of children is postponed until the child can participate in the decision. In response, the Committee on Bioethics intends to organise a seminar focusing on how the Resolution can be upheld in practice, by identifying good practices in dealing with interventions on intersex children.

Safeguarding the rights of persons with mental health difficulties.

25. The issue of mental health is expected to be one of the biggest challenges facing healthcare systems in the future. Mental healthcare should be treated no differently to physical healthcare in that a human rights-based approach should be adopted in both. It is vital that the rights and self-determination of all patients, including persons with mental health difficulties, be promoted and that they may actively participate to the greatest possible extent in all decisions regarding their treatment and care. In this context, the development and use of voluntary measures and practices in mental healthcare should be promoted.

Actions

- **Elaborating a legal instrument to protect the human rights and dignity of persons with mental disorders with regard to involuntary placement and/or involuntary treatment.**

  The deprivation of liberty involved in involuntary placement and treatment impacts on a person's right to freedom from cruel, inhuman or degrading treatment (Article 3), right to liberty (Article 5), and the right to respect for private life (Article 8) as enshrined in the European Convention on Human Rights. In this connection, Article 5 of the Oviedo Convention refers to the principle of free and informed consent for any medical treatment. Article 7 of the Oviedo Convention constitutes an exception to the general rule of consent for the protection of persons who have a mental disorder. To this end, three conditions must be satisfied: the person must have a serious mental health problem; the treatment must aim to alleviate the mental health problem; and without treatment of the mental health problem, serious harm to their health is likely to result. More recently, Recommendation Rec(2004)10 of the Committee of Ministers has detailed the conditions under which a person may be subjected to compulsory medical treatment (Article 18) and the conditions for involuntary treatment (Article 19). The Committee on Bioethics seeks to build on its previous work in this area to ensure that involuntary detention of persons is a last resort and, in this case, when strictly necessary, that the human rights and dignity of patients are consistently and effectively upheld.

- **Developing a Compendium of good practices to promote voluntary measures in the field of mental healthcare.**

  In mental healthcare for persons with psychosocial disabilities the focus is shifting towards avoiding recourse to involuntary measures. To assist member States in this shift, the Committee on Bioethics intends to develop a Compendium of good practices to promote voluntary measures in mental healthcare, both at a preventive level and in situations of crisis, by focusing on examples in member States.

CO-OPERATION AND COMMUNICATION

26. Many of the challenges raised by new developments in biomedicine necessitate effective and efficient co-operation with other organisations and bodies. This is an opportunity to share knowledge, experience, and skills. It also allows for the pursuit of mutual interests and the realisation of common goals in innovative ways, with synergy and without duplication of resources. The importance and relevance of such co-operation is reflected in the objective of the UN Interagency Committee on Bioethics to which the Council of Europe is an associate member. Co-operation concerns both normative and methodological aspects, i.e. how and on what issues the Committee on Bioethics should co-operate with other actors in the field. All actions should be visible, and achievements strategically communicated to raise awareness and to inform public policy. Consequently, the fourth pillar of the Strategic Action Plan is focused on transversal co-operation and communication as a prerequisite for achieving the strategic objectives in the Strategic Action Plan.

Developing long-term strategic co-operation with Council of Europe committees and other intergovernmental bodies working in the field of bioethics.

27. The resources of the Committee on Bioethics should be deployed to maximise its efficiency and to ensure that it makes a unique contribution to the challenges presented to it. It is therefore essential for the Committee to develop long-term strategic co-operation with other actors in the field of bioethics, both within and outside of the Council of Europe.

Actions

- **Reviewing the working methods of the Committee on Bioethics in order to elaborate a Framework for effective co-operation with Council of Europe committees and other intergovernmental organisations working in the field of bioethics.**

  In view of the need to ensure effective co-operation, the Committee on Bioethics considers it important to review its working methods and to elaborate a standardised mechanism for co-operation with other bodies. A framework should set out ways to prioritise requests to comment on initiatives from other bodies and to strengthen collaboration with the Parliamentary Assembly of the Council of Europe, other Council of Europe committees, other intergovernmental organisations working in the field of bioethics, and policy makers at member State level, in order to best achieve shared objectives.

- **Establishing links and co-operation with National Training Institutions to help disseminate the HELP course on bioethics in Council of Europe member States.**

  The European Programme for Human Rights Education for Legal Professionals (HELP), together with the Bioethics Unit of the Council of Europe, have developed an online training course on bioethics. The course addresses ethical and legal issues raised by developments in the field of biomedicine and brings attention to

the principles enshrined in legal instruments developed by the Committee on Bioethics and other committees and bodies of the Council of Europe, and adopted by the Committee of Ministers, as well as to relevant case law of the European Court of Human Rights. To raise awareness of key human rights principles in the biomedical field and to encourage interdisciplinary interaction and learning, the Committee on Bioethics intends to disseminate its HELP course on bioethics not only to legal professionals but also to health professionals and other categories of users. This includes its roll-out in cooperation with the National Training Institutions for legal and health professionals.

Ensuring the communication and dissemination of the outputs of the Committee on Bioethics to internal and external stakeholders in order to maximise their uptake and utility.

28. To raise awareness of human rights principles and the challenges raised by developments in the field of biomedicine, it is important for the work of the Committee on Bioethics to be widely communicated and rendered more visible to all stakeholders. This will facilitate an increased understanding of the contribution of the Committee on Bioethics, and of the Council of Europe more generally, to protecting human rights in the field of biomedicine. It is therefore essential for the Committee on Bioethics to develop effective dissemination strategies for its outputs, which are accessible to a wide range of different relevant stakeholders. This helps to inform public policy. This will require considering the most effective ways to communicate outputs to target audiences and to engage stakeholders throughout the process. In this regard, it is important to recognise that young people should be a key focal point for bioethical deliberations, considering that they will experience the impacts of emerging and converging technologies and that they will be shaping the future of society.

Actions

- **Developing an annual online newsletter covering the work of the Committee on Bioethics, bioethical developments in Council of Europe member States, and the case law of the European Court of Human Rights.**

  To ensure communication and dissemination of bioethical developments in the Council of Europe, an annual online newsletter oriented towards lay audiences should be developed. The newsletter should provide information on the work of the Committee on Bioethics and its impact on Council of Europe member States, the work of committees and other bodies of the Council of Europe in the field of biomedicine, relevant case law of the European Court of Human Rights, and bioethical developments in Council of Europe member States. It intends to serve as a platform for information to be shared between member States, making connections between States with similar interests. It should also serve as a useful means of communicating and promoting the work of the Committee on Bioethics and the Council of Europe to relevant third parties.

- **Hosting a Youth Forum for Bioethics to provide young people with an opportunity to share their views on bioethical topics and to inform the work of the Committee on Bioethics.**

As a way to bring the voice of European youth into bioethics discussions at the Council of Europe, the Committee on Bioethics intends to host a (one-off) Youth Forum. The Forum should provide a space for younger people to interact with the Committee on Bioethics and to provide input on bioethical issues, thereby empowering them to represent and advocate their needs and interests. Moreover, it will provide the Committee on Bioethics with valuable insights from younger people to inform its own work. The Youth Forum should, where appropriate, act as a model that could be used in the future.

IMPLEMENTATION

Timeframe: The Strategic Action Plan is intended to be implemented within a timeframe of six years (2020-2025).

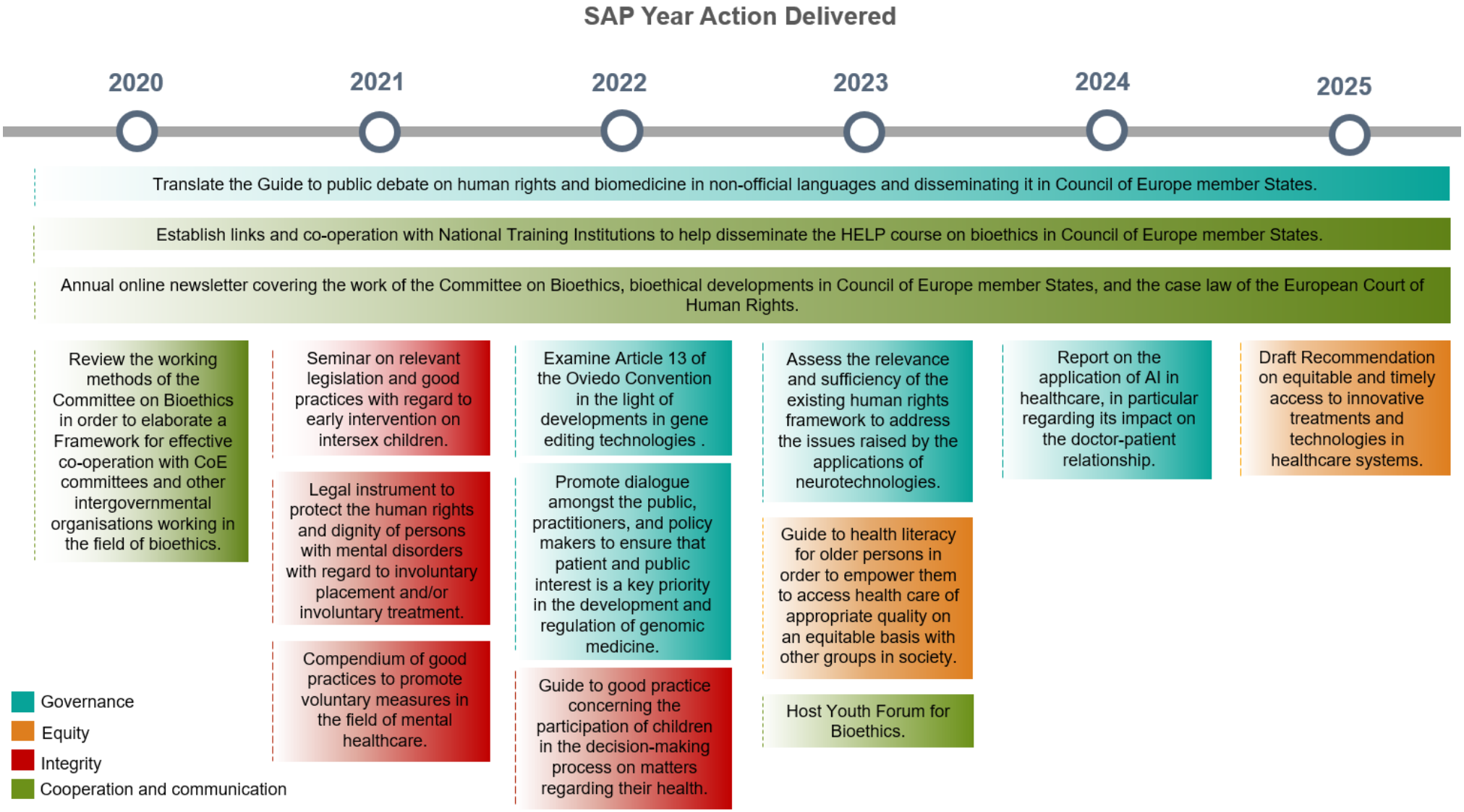
Methodology: The proposed actions will be carried out in the light of the principles set out in the Oviedo Convention and its Additional Protocols, in the relevant Recommendations of the Committee of Ministers, as well as in reports, guides, and position statements issued by the Committee on Bioethics. The findings of the International Conference on the 20th Anniversary of the Oviedo Convention, the International Conference on Emerging Technologies and Human Rights, and the High-Level Seminar on International Case-Law in Bioethics will also be used as a basis. The proposed actions take into account complementarity and co-operation with other Council of Europe bodies and other relevant intergovernmental organisations.

Gender equality and diversity: Throughout the implementation of the Strategic Action Plan, gender equality and respect for diversity will be ensured, in particular in partnership with the Gender Equality Rapporteur designated by the Committee on Bioethics. Gender balance and respect for diversity will be promoted in the composition of working groups and panels and in the appointment of rapporteurs, chairs, and external experts. An approach that is sensitive to gender equality and diversity has been integrated in the process of identifying priorities for the Strategic Action Plan and gender equality and diversity-specific challenges that may arise during the implementation of its actions will be monitored, evaluated, and addressed.

Leadership: The actions proposed under the Strategic Action Plan are intended to be carried out under the responsibility of the Committee on Bioethics and, where appropriate, in co-ordination with other Council of Europe bodies or intergovernmental organisations.

Funding: The implementation of actions will be covered by existing budgetary allocations provided from the Council of Europe's Ordinary Budget. For some actions, such as translating and disseminating the Guide to public debate on human rights and biomedicine, the roll-out of the HELP course on bioethics, and the DH-BIO Youth Forum, the implementation depends on voluntary contributions.

Reporting: The Committee on Bioethics will prepare mid-term and final reports to be communicated to the Steering Committee on Human Rights and to the Committee of Ministers. The mid-term report will contain a review of progress in respect of the objectives and actions in the Strategic Action Plan, and an assessment of their ongoing relevance.



The timeline indicates the year in which it is estimated that the action will be delivered; work will be initiated in advance of the year indicated as there may be a number of sub-actions required to realise the final outcome.

LIST OF SOURCES

- Proceedings of the International Conference on the 20th Anniversary of the Oviedo Convention: Relevance and Challenges, 24-25 October 2017, Strasbourg, available at https://rm.coe.int/english-proceedings-20-anni/168089e570

- Rapporteur Report on the International Conference on the 20th Anniversary of the Oviedo Convention: Relevance and Challenges, 24-25 October 2017, Strasbourg, available at https://rm.coe.int/oviedo-conference-rapporteur-report-e/168078295c

- The Rights of Children in Biomedicine: Challenges Posed by Scientific Advances and Uncertainties. Report prepared by Zillén, Kavot; Garland, Jameson and Slokenberga, Santa. Uppsala University, 2017, available at https://rm.coe.int/16806d8e2f

- From Law to Practice: Towards a Roadmap to Strengthen Children's Rights in the Era of Biomedicine. Report prepared by Liefaard, Ton; Hendriks, Aart and Zlotnik, Daniella. Universiteit Leiden, 2017, available at https://rm.coe.int/leiden-university-report-biomedicine-final/168072fb46

- Proceedings of the High Level Seminar on International Case-Law in Bioethics: Insight and Foresight, 5 December 2016, Strasbourg, available at https://rm.coe.int/proceedings-caselaw-/1680736452

- Proceedings of the International Conference on Emerging Technologies and Human Rights, 4-5 May 2015, Strasbourg, available at https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680495b44

- Report on Ethical Issues Raised by Emerging Sciences and Technologies. Report prepared by Strand, Roger and Kaiser, Matthias. Bergen University, 2015, available at https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168030751d

- From Bio to NBIC Convergence – From Medical Practice to Daily Life. Report prepared by van Est, Rinie; Stemerding, Dirk; Rerimassie, Virgil; Schuijff, Mirjam; Timmer, Jelte and Brom, Frans. The Hague, Rathenau Instituut, 2014, available at https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680307575

International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(2 June 2022)

Focus Group on Artificial Intelligence for Health (FG-AI4H)

FG-AI4H DEL01

Ethics and governance of artificial intelligence for health

Summary

Artificial Intelligence (AI) refers to the ability of algorithms encoded in technology to learn from data so that they can perform automated tasks without every step in the process having to be programmed explicitly by a human. While AI holds great promise for the practice of public health and medicine, ethical challenges for health care systems, practitioners and beneficiaries of medical and public health services must be addressed. Many of the ethical concerns described in this document predate the advent of AI, although AI itself presents a number of novel concerns.

This document endorses a set of six key ethical principles:

– Protect human autonomy
– Promote human well-being and safety and the public interest
– Ensure transparency, explainability and intelligibility
– Foster responsibility and accountability
– Ensure inclusiveness and equity
– Promote AI that is responsive and sustainable

It is hoped that these principles will be used as a basis for governments, technology developers, companies, civil society and inter-governmental organizations to adopt ethical approaches to appropriate use of AI for health.

Keywords

Artificial intelligence; AI for health; Ethics; Governance

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1 of the Deliverable FG-AI4H DEL01 on "*Ethics and governance of artificial intelligence for health*" [approved at the ITU-T Focus Group on AI for Health (FG-AI4H) meeting held in (2 June 2022)].

Editors: Andreas Reis, World Health Organization, E-mail: reisa@who.int

Sameer Pujari, World Health Organization, E-mail: pujaris@who.int

© ITU 2022

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Acknowledgements

Original development of this guidance document was led by Andreas Reis (Co-Lead, Health Ethics and Governance Unit, department of Research for Health) and Sameer Pujari (department of Digital Health and Innovation), under the overall guidance of John Reeder (Director, Research for Health), Bernardo Mariano (Director, Digital Health and Innovation) and Soumya Swaminathan (Chief Scientist).

Rohit Malpani (consultant, France) was the lead writer. The Co-Chairs of the Expert Group, Effy Vayena (ETH Zurich, Switzerland) and Partha Majumder (National Institute of Biomedical Genomics, India), provided overall guidance for the drafting of this document.

WHO is grateful to the following individuals who contributed to development of this guidance.

External expert group

Najeeb Al Shorbaji, eHealth Development Association, Jordan

Arisa Ema, Ito International Research Center (Institute for Future Initiative), Japan

Amel Ghoulia, H3Africa, H3ABioNet, Tunisia

Jennifer Gibson, Joint Centre for Bioethics, Dalla Lana School of Public Health, University of Toronto, Canada

Kenneth W. Goodman, Institute for Bioethics and Health Policy, University of Miami Miller School of Medicines, USA

Jeroen van den Hoven, Delft University of Technology, The Netherlands

Malavika Jayaram, Digital Asia Hub, Singapore

Daudi Jjingo, Makerere University, Uganda

Tze Yun Leong, National University of Singapore, Singapore.

Alex John London, Carnegie Mellon University, USA

Partha Majumder, National Institute of Biomedical Genomics, India

Tshilidzi Marwala, University of Johannesburg, South Africa

Roli Mathur, Indian Council of Medical Research, India

Timo Minssen, Centre for Advanced Studies in Biomedical Innovation Law (CeBIL), Faculty of Law, University of Copenhagen, Denmark

Andrew Morris, Health Data Research UK, United Kingdom

Daniela Paolotti, ISI Foundation, Italy

Maria Paz Canales, Derechos Digitales, Chile

Jerome Singh, University of Kwa-Zulu Natal, South Africa

Effy Vayena, ETH Zurich, Switzerland

Robyn Whittaker, University of Auckland, New Zealand

Yi Zeng, Chinese Academy of Sciences, China

Observers

Tee Wee Ang, United Nations Educational, Scientific and Cultural Organization, France

Abdoulaye Banire Diallo, University of Quebec at Montreal, Canada

Julien Durand, Takeda, Switzerland

David Gruson, Jouve, France

Lee Hibbard, Council of Europe, France

Lauren Milner, US Food and Drug Administration, USA

Rasha Abdul Rahim, Amnesty Tech, United Kingdom

Elettra Ronchi, Organization for Economic Co-operation and Development, France

External reviewers

Anurag Aggarwal, Council of Scientific and Industrial Research, India
Paolo Alcini, European Medicines Agency, Netherlands
Pamela Andanda, University of Witwatersrand, South Africa
Eva Blum-Dumontet, Privacy International, United Kingdom
Marcelo Corrales Compagnucci, CeBIL, Faculty of Law, University of Copenhagen, Denmark
Sara Leila Meg Davis, Graduate Institute, Switzerland
Juan M. Duran, Delft University of Technology, Netherlands
Osama El-Hassan, Dubai Health Authority, United Arab Emirates
Tomaso Falchetta, Privacy International, United Kingdom
Sara Gerke, Harvard Law School, USA
Tabitha Ha, STOP AIDS, United Kingdom
Henry Hoffman, ADA Health, Germany
Calvin Ho, University of Hong Kong, Hong Kong (China)
Prageeth Jayathissa, Vector Ltd, New Zealand
Otmar Kloiber, World Medical Association, Switzerland
Paulette Lacroix, International Medical Informatics Association, Canada
Hannah Lim, National University of Singapore, Singapore
Allan Maleche, Kenya Legal and Ethical Issues Network on HIV and AIDS, Kenya
Peter Micek, Access Now, USA
Thomas Neumark, University of Oslo, Norway
Laura O'Brien, Access Now, USA
Alexandrine Pirlot de Corbion, Privacy International, United Kingdom
Léonard Van Rompaey, University of Copenhagen, Denmark
Tony Joakim Sandset, University of Oslo, Norway
Jay Shaw, Women's College Hospital Institute for Health System Solutions and Virtual Care, Canada
Sam Smith, medConfidential, United Kingdom
David Stewart, International Council of Nurses, Switzerland

External presenters at expert meetings

David Barbe, World Medical Association, USA
Elisabeth Bohn, Academy of Medical Sciences, United Kingdom
Katherine Chou, Google, USA
I. Glenn Cohen, Harvard Law School, USA
Naomi Lee, The Lancet, United Kingdom
Nada Malou, Médecins Sans Frontières, France
Vasantha Muthuswamy, Indian Council of Medical Research (retired), India
Sharon Kaur, A/P Gurmukh Singh, University of Malaya, Malaysia
Christian Stammel, Wearable Technologies, Germany
Alex Wang, Tencent, China
Kirstie Whitaker, Turing Institute, United Kingdom
Thomas Wiegand, Fraunhofer Heinrich Hertz Institute, Germany

WHO staff

Onyema Ajuebor, Technical Officer, Health Workforce, Geneva

Shada Al-Salamah, Consultant, Digital Health and Innovation, Geneva

Ryan Dimentberg, Intern, Health Ethics and Governance Unit, Geneva

Clayton Hamilton, Technical Officer, WHO Regional Office for Europe, Copenhagen

Katherine Littler, Co-Lead, Health Ethics and Governance Unit, Geneva

Rohit Malpani, Consultant, Health Ethics and Governance Unit, Geneva

Ahmed Mohamed Amin Mandil, Coordinator, Research and Innovation, WHO Regional Office for the Eastern Mediterranean, Cairo

Bernardo Mariano, Chief Information Officer, Geneva

Issa T. Matta, Legal Affairs, Geneva

Vasee Moorthy, Coordinator, Health Systems and Innovation, Information, Evidence and Research, Research, Ethics and Knowledge Management, Geneva

Mohammed Hassan Nour, Technical Officer, Digital Health and Innovation, WHO Regional Office for the Eastern Mediterranean, Cairo

Lee-Anne Pascoe, Consultant, Health Ethics and Governance Unit, Geneva

Sameer Pujari, Technical Officer, Digital Health and Innovation, Geneva

Andreas Reis, Co-Lead, Health Ethics and Governance Unit, Geneva

Soumya Swaminathan, Chief Scientist, Geneva

Mariam Shokralla, Consultant, Digital Health and Innovation, Geneva

Diana Zandi, Technical Officer, Integrated Health Services, Geneva

Yu Zhao, Technical Officer, Digital Health and Innovation, Geneva

Table of Contents

		Page
Executive summary		x
1	Introduction	1
2	Artificial intelligence	3
3	Applications of artificial intelligence for health	4
3.1	In health care	4
3.1.1	Diagnosis and prediction-based diagnosis	4
3.1.2	Clinical care	5
3.1.3	Emerging trends in the use of AI in clinical care	5
3.2	In health research and drug development	7
3.2.1	Application of AI for health research	7
3.2.2	Uses of AI in drug development	7
3.3	In health systems management and planning	8
3.4	In public health and public health surveillance	8
3.4.1	Health promotion	8
3.4.2	Disease prevention	9
3.4.3	Surveillance (including prediction-based surveillance) and emergency preparedness	9
3.4.4	Outbreak response	10
3.5	The future of artificial intelligence for health	10
4	Laws, policies and principles that apply to artificial intelligence for health	11
4.1	Artificial intelligence and human rights	11
4.2	Data protection laws and policies	12
4.3	Existing laws and policies related to health data	12
4.4	General principles for the development and use of artificial intelligence	13
4.5	Principles for use of artificial intelligence for health	14
4.6	Bioethics laws and policies	14
4.7	Regulatory considerations	14
5	Key ethical principles for use of artificial intelligence for health	15
5.1	Protect autonomy	16
5.2	Promote human well-being, human safety and the public interest.	17
5.3	Ensure transparency, explainability and intelligibility.	17
5.4	Foster responsibility and accountability.	18
5.5	Ensure inclusiveness and equity.	19
5.6	Promote artificial intelligence that is responsive and sustainable.	20
6	Ethical challenges to use of artificial intelligence for health care	20
6.1	Assessing whether artificial intelligence should be used	20
6.2	Artificial intelligence and the digital divide	22
6.3	Data collection and use	23

Page

6.3.1 Data colonialism 27
6.3.2 Mechanisms for safeguarding privacy – Do they work? 27
6.4 Accountability and responsibility for decision-making with artificial intelligence 28
6.4.1 Accountability for AI-related errors and harm 30
6.5 Autonomous decision-making 31
6.5.1 Implications of replacing human judgement for clinical care 31
6.5.2 Implications of the loss of human control in clinical care 32
6.5.3 The ethics of using AI for resource allocation and prioritization 34
6.5.4 Use of AI for predictive analytics in health care 34
6.5.5 Use of AI for prediction in drug discovery and clinical development 36
6.6 Bias and discrimination associated with artificial intelligence 37
6.6.1 Bias in data 37
6.6.2 Biases related to who develops AI and the origin of the data on which AI is trained 38
6.6.3 Bias in deployment 39
6.7 Risks of artificial intelligence technologies to safety and cybersecurity 39
6.7.1 Safety of AI technologies 39
6.7.2 Cybersecurity 39
6.8 Impacts of artificial intelligence on labour and employment in health and medicine 40
6.9 Challenges in commercialization of artificial intelligence for health care 42
6.10 Artificial intelligence and climate change 43
7 Building an ethical approach to use of artificial intelligence for health 44
7.1 Ethical, transparent design of technologies 44
Putting prediction to good use 46
7.2 Engagement and role of the public and demonstration of trustworthiness to providers and patients 47
7.3 Impact assessment 49
7.4 Research agenda for ethical use of artificial intelligence for health care 50
8 Liability regimes for artificial intelligence for health 51
8.1 Liability for use of artificial intelligence in clinical care 51
8.2 Are machine-learning algorithms products? 52
8.3 Compensation for errors 53
8.4 Role of regulatory agencies and pre-emption 53
8.5 Considerations for low- and middle-income countries 53
9 Elements of a framework for governance of artificial intelligence for health 54
9.1 Governance of data 55
9.1.1 Evolving approaches to consent 55

Page

9.1.2 Broad consent ... 56
9.1.3 Data protection ... 57
9.1.4 Community control of health data – data sovereignty and data cooperatives ... 58
9.1.5 Federated data ... 58
9.1.6 Government principles and guidelines ... 58
9.1.7 Data-sharing, including data hubs ... 60
9.1.8 Data hubs ... 60
9.1.9 Data-sharing and data partnerships with the private sector ... 60
9.2 Control and benefit-sharing ... 61
9.2.1 Control over and benefit-sharing of big data ... 61
9.2.2 Ownership of AI-based products, services and methods ... 63
9.3 Governance of the private sector ... 64
9.3.1 The role of self-governance ... 65
9.3.2 Public-private partnerships for AI for health care ... 66
9.3.3 Governance and oversight of large technology companies ... 67
9.3.4 Provision of health care by the private sector outside the health-care system ... 68
9.3.5 An enabling environment for effective governance of the private sector ... 68
9.4 Governance of the public sector ... 69
9.4.1 Assessing whether AI is necessary and appropriate for use in the public sector ... 70
9.4.2 Accountability through transparency and participation ... 70
9.4.3 Appropriate collection, stewardship and use of data ... 70
9.4.4 Risks and opportunities in use of AI for provision of public services and social protection ... 71
9.5 Regulatory considerations ... 72
9.5.1 Does regulation stifle innovation? ... 72
9.5.2 Transparency and explainability of AI-based devices ... 72
9.5.3 Addressing bias ... 74
9.5.4 Ethical considerations for LMIC and HIC with poor health outcomes ... 74
9.6 Policy observatory and model legislation ... 75
9.7 Global governance of artificial intelligence ... 75
References ... 78
Annex A – Considerations for the ethical design, deployment and use of artificial intelligence technologies for health ... 102
A.1 Considerations for AI developers ... 102
A.1.1 Designing an AI technology ... 102
A.1.2 Developing an AI technology ... 104

Page
A.1.3 Deploying an AI technology and improving it after deployment 105
A.2 Considerations for ministries of health..... 106
A.2.1 How to protect the health and safety of patients 106
A.2.2 Prepare for the introduction and use of AI technologies..... 107
A.2.3 Address ethical and legal challenges and protect human rights..... 108
A.3 Considerations for health-care institutions and providers 110
A.3.1 Is the AI technology necessary and appropriate?..... 110
A.3.2 Is the context in which the AI technology will be used appropriate? 111
A.3.3 Should a health-care provider use the AI technology? 112

List of Boxes

Box 1 – Examples of AI ethics principles proposed by intergovernmental organizations and countries..... 13
Box 2 – The emergence of digital identification in the COVID-19 pandemic 25
Box 3 – Dinerstein vs Google 26
Box 4 – Informed consent during clinical care 33
Box 5 – Challenges associated with a system for predicting adolescent pregnancy in Argentina 36
Box 6 – Discrimination and racial bias in AI technology..... 37
Box 7 – AI technologies for detecting skin cancer exclude people of colour. 38
Box 8 – Design for values [229] 45
Box 9 – Supporting health workers in the use AI technologies, including through education and training 48

List of Figures

Figure 1 – Health data ecosystem [115]..... 24
Figure 2 – Elements of transparent data use [280]..... 59

Abbreviations and acronyms

AI	artificial intelligence
CeBIL	Centre for Advanced Studies in Biomedical Innovation Law
EU	European Union
GDPR	General Data Protection Regulation
HIC	high-income countries
IP	intellectual property
LMIC	low- and middle-income countries
NHS	National Health Service (United Kingdom)
OECD	Organization for Economic Co-operation and Development
PPP	private-public partnership
SOFA	Sequential Organ Failure Assessment
UNESCO	United Nations Economic, Scientific and Cultural Organization
US	United States (of America)
USA	United States of America

Executive summary

Artificial Intelligence (AI) refers to the ability of algorithms encoded in technology to learn from data so that they can perform automated tasks without every step in the process having to be programmed explicitly by a human. WHO recognizes that AI holds great promise for the practice of public health and medicine. WHO also recognizes that, to fully reap the benefits of AI, ethical challenges for health care systems, practitioners and beneficiaries of medical and public health services must be addressed. Many of the ethical concerns described in this document predate the advent of AI, although AI itself presents a number of novel concerns.

Whether AI can advance the interests of patients and communities depends on a collective effort to design and implement ethically defensible laws and policies and ethically designed AI technologies. There are also potential serious negative consequences if ethical principles and human rights obligations are not prioritized by those who fund, design, regulate or use AI technologies for health. AI's opportunities and challenges are thus inextricably linked.

AI can augment the ability of health-care providers to improve patient care, provide accurate diagnoses, optimize treatment plans, support pandemic preparedness and response, inform the decisions of health policy-makers or allocate resources within health systems. To unlock this potential, health-care workers and health systems must have detailed information on the contexts in which such systems can function safely and effectively, the conditions necessary to ensure reliable, appropriate use, and the mechanisms for continuous auditing and assessment of system performance. Health-care workers and health systems must have access to education and training in order to use and maintain these systems under the conditions for their safe, effective use.

AI can also empower patients and communities to assume control of their own health care and better understand their evolving needs. To achieve this, patients and communities require assurance that their rights and interests will not be subordinated to the powerful commercial interests of technology companies or the interests of governments in surveillance and social control. It also requires that the potential of AI to detect risks to patient or community health is incorporated into health systems in a way that advances human autonomy and dignity and does not displace humans from the centre of health decision-making.

AI can enable resource-poor countries, where patients often have restricted access to health-care workers or medical professionals, to bridge gaps in access to health services. AI systems must be carefully designed to reflect the diversity of socio-economic and health-care settings and be accompanied by training in digital skills, community engagement and awareness-raising. Systems based primarily on data of individuals in high-income countries may not perform well for individuals in low- and middle-income settings. Country investments in AI and the supporting infrastructure should therefore help to build effective health-care systems by avoiding AI that encodes biases that are detrimental to equitable provision of and access to health-care services.

This publication was issued in 2021 by the WHO as a WHO Guidance [337] and approved by the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) as its Deliverable 1 at its Meeting O in Berlin, 31 May – 2 June 2022. It was originally produced jointly by WHO's Health Ethics and Governance unit in the department of Research for Health and by the department of Digital Health and Innovation, is based on the collective views of a WHO Expert Group on Ethics and Governance of AI for Health, which comprised 20 experts in public health, medicine, law, human rights, technology and ethics. FG-AI4H experts also contributed to the preparation of the document. The group analysed many opportunities and challenges of AI and recommended policies, principles and practices for ethical use of AI for health and means to avoid its misuse to undermine human rights and legal obligations.

AI for health has been affected by the COVID-19 pandemic. Although the pandemic is not a focus of this document, it has illustrated the opportunities and challenges associated with AI for health. Numerous new applications have emerged for responding to the pandemic, while other applications

have been found to be ineffective. Several applications have raised ethical concerns in relation to surveillance, infringement on the rights of privacy and autonomy, health and social inequity and the conditions necessary for trust and legitimate uses of data-intensive applications. During their deliberations on this document, members of the expert group prepared interim WHO guidance for the use of proximity tracking applications for COVID-19 contact-tracing.

Key ethical principles for the use of AI for health

This document endorses a set of key ethical principles. It is hoped that these principles will be used as a basis for governments, technology developers, companies, civil society and inter-governmental organizations to adopt ethical approaches to appropriate use of AI for health. The six principles are summarized below and explained in depth in section 5.

Protecting human autonomy: Use of AI can lead to situations in which decision-making power could be transferred to machines. The principle of autonomy requires that the use of AI or other computational systems does not undermine human autonomy. In the context of health care, this means that humans should remain in control of health-care systems and medical decisions. Respect for human autonomy also entails related duties to ensure that providers have the information necessary to make safe, effective use of AI systems and that people understand the role that such systems play in their care. It also requires protection of privacy and confidentiality and obtaining valid informed consent through appropriate legal frameworks for data protection.

Promoting human well-being and safety and the public interest. AI technologies should not harm people. The designers of AI technologies should satisfy regulatory requirements for safety, accuracy and efficacy for well-defined use cases or indications. Measures of quality control in practice and quality improvement in the use of AI over time should be available. Preventing harm requires that AI not result in mental or physical harm that could be avoided by use of an alternative practice or approach.

Ensuring transparency, explainability and intelligibility. AI technologies should be intelligible or understandable to developers, medical professionals, patients, users and regulators. Two broad approaches to intelligibility are to improve the transparency of AI technology and to make AI technology explainable. Transparency requires that sufficient information be published or documented before the design or deployment of an AI technology and that such information facilitate meaningful public consultation and debate on how the technology is designed and how it should or should not be used. AI technologies should be explainable according to the capacity of those to whom they are explained.

Fostering responsibility and accountability. Humans require clear, transparent specification of the tasks that systems can perform and the conditions under which they can achieve the desired performance. Although AI technologies perform specific tasks, it is the responsibility of stakeholders to ensure that they can perform those tasks and that AI is used under appropriate conditions and by appropriately trained people. Responsibility can be assured by application of "human warranty", which implies evaluation by patients and clinicians in the development and deployment of AI technologies. Human warranty requires application of regulatory principles upstream and downstream of the algorithm by establishing points of human supervision. If something goes wrong with an AI technology, there should be accountability. Appropriate mechanisms should be available for questioning and for redress for individuals and groups that are adversely affected by decisions based on algorithms.

Ensuring inclusiveness and equity. Inclusiveness requires that AI for health be designed to encourage the widest possible appropriate, equitable use and access, irrespective of age, sex, gender, income, race, ethnicity, sexual orientation, ability or other characteristics protected under human rights codes. AI technology, like any other technology, should be shared as widely as possible. AI technologies should be available for use not only in contexts and for needs in high-income settings but also in the contexts and for the capacity and diversity of LMIC. AI technologies should not encode

biases to the disadvantage of identifiable groups, especially groups that are already marginalized. Bias is a threat to inclusiveness and equity, as it can result in a departure, often arbitrary, from equal treatment. AI technologies should minimize inevitable disparities in power that arise between providers and patients, between policy-makers and people and between companies and governments that create and deploy AI technologies and those that use or rely on them. AI tools and systems should be monitored and evaluated to identify disproportionate effects on specific groups of people. No technology, AI or otherwise, should sustain or worsen existing forms of bias and discrimination.

Promoting AI that is responsive and sustainable. Responsiveness requires that designers, developers and users continuously, systematically and transparently assess AI applications during actual use. They should determine whether AI responds adequately and appropriately and according to communicated, legitimate expectations and requirements. Responsiveness also requires that AI technologies be consistent with wider promotion of the sustainability of health systems, environments and workplaces. AI systems should be designed to minimize their environmental consequences and increase energy efficiency. That is, use of AI should be consistent with global efforts to reduce the impact of human beings on the Earth's environment, ecosystems and climate. Sustainability also requires governments and companies to address anticipated disruptions in the workplace, including training for health-care workers to adapt to the use of AI systems, and potential job losses due to use of automated systems.

Overview of this document

This document is divided into nine sections and an annex. Section 1 explains the rationale for WHO's engagement in this topic and the intended readership of the document's findings, analyses and recommendations. Sections 2 and 3 define AI for health through its methods and applications. Section 2 provides a non-technical definition of AI, which includes several forms of machine learning as a subset of AI techniques. It also defines "big data", including sources of data that comprise biomedical or health big data. Section 3 provides a non-comprehensive classification and examples of AI technologies for health, including applications used in LMIC, such as for medicine, health research, drug development, health systems management and planning, and public health surveillance.

Section 4 summarizes the laws, policies and principles that apply or could apply to the use of AI for health. These include human rights obligations as they apply to AI, the role of data protection laws and frameworks and other health data laws and policies. The section describes several frameworks that commend ethical principles for the use of AI for health, as well as the roles of bioethics, law, public policy and regulatory frameworks as sources of ethical norms.

Section 5 describes the six ethical principles that the Expert Group identified as guiding the development and use of AI for health. Section 6 presents the ethical challenges identified and discussed by the Expert Group to which these guiding ethical principles can be applied: whether AI should be used; AI and the digital divide; data collection and use; accountability and responsibility for decision-making with AI; autonomous decision-making; bias and discrimination associated with AI; risks of AI to safety and cybersecurity; impacts of AI on labour and employment in health care; challenges in the commercialization of AI for health care; and AI and climate change.

The final sections of the document identify legal, regulatory and non-legal measures for promoting ethical use of AI for health, including appropriate governance frameworks. Recommendations are provided.

Section 7 examines how various stakeholders can introduce ethical practices, programmes and measures to anticipate or meet ethical norms and legal obligations. They include ethical, transparent design of AI technologies; mechanisms for the engagement and role of the public and demonstrating trustworthiness with providers and patients; impact assessment; and a research agenda for ethical use of AI for health care.

Section 8 is a discussion of how liability regimes may evolve with increasing use of AI for health care. It includes how liability could be assigned to a health-care provider, a technology provider and a health-care system or hospital that selects an AI technology and how the rules of liability might influence how a practitioner uses AI. The section also considers whether machine-learning algorithms are products, how to compensate individuals harmed by AI technologies, the role of regulatory agencies and specific aspects for LMIC.

Section 9 presents elements of a governance framework for AI for health. "Governance in health" refers to a range of functions for steering and rule-making by governments and other decision-makers, including international health agencies, to achieve national health policy objectives conducive to universal health coverage. The section analyses several governance frameworks either being developed or already matured. The frameworks discussed are: governance of data, control and benefit-sharing, governance of the private sector, governance of the public sector, regulatory considerations, the role of a policy observatory and model legislation and global governance of AI.

Finally, the document provides practical advice for implementing the WHO guidance for three sets of stakeholders: AI technology developers, ministries of health and health-care providers. The considerations are intended only as a starting-point for context-specific discussions and decisions by diverse stakeholders.

While the primary readership of this guidance document is ministries of health, it is also intended for other government agencies, ministries that will regulate AI and those who use AI technologies for health. The guidance is also intended for entities that design and finance AI technologies for health.

Implementation of this guidance will require collective action. Companies and governments should introduce AI technologies only to improve the human condition and not for objectives such as unwarranted surveillance or to increase the sale of unrelated commercial goods and services. Providers should demand appropriate technologies and use them to maximize both the promise of AI and clinicians' expertise. Patients, community organizations and civil society should be able to hold governments and companies to account, to participate in the design of technologies and rules, to develop new standards and approaches and to demand and seek transparency to meet their own needs as well as those of their communities and health systems.

AI for health is a fast-moving, evolving field, and many applications, not yet envisaged, will emerge with ever-greater public and private investment. WHO may consider issuing specific guidance for additional tools and applications and may update this guidance periodically to keep pace with this rapidly changing field.

ITU-T FG-AI4H Deliverable FG-AI4H DEL01

Ethics and governance of artificial intelligence for health

1 Introduction

Digital technologies and artificial intelligence (AI), particularly machine learning, are transforming medicine, medical research and public health. Technologies based on AI are now used in health services in countries of the Organization for Economic Co-operation and Development (OECD), and its utility is being assessed in low- and middle-income countries (LMIC). The United Nations Secretary-General has stated that safe deployment of new technologies, including AI, can help the world to achieve the United Nations Sustainable Development Goals [1], which would include the health-related objectives under Sustainable Development Goal 3. AI could also help to meet global commitments to achieve universal health coverage.

Use of AI for health nevertheless raises trans-national ethical, legal, commercial and social concerns. Many of these concerns are not unique to AI. The use of software and computing in health care has challenged developers, governments and providers for half a century, and AI poses additional, novel ethical challenges that extend beyond the purview of traditional regulators and participants in health-care systems. These ethical challenges must be adequately addressed if AI is to be widely used to improve human health, to preserve human autonomy and to ensure equitable access to such technologies.

Use of AI technologies for health holds great promise and has already contributed to important advances in fields such as drug discovery, genomics, radiology, pathology and prevention. AI could assist health-care providers in avoiding errors and allow clinicians to focus on providing care and solving complex cases. The potential benefits of these technologies and the economic and commercial potential of AI for health care presage ever greater use of AI worldwide.

Unchecked optimism in the potential benefits of AI could, however, veer towards habitual first recourse to technological solutions to complex problems. Such "techno-optimism" could make matters worse, for example, by exacerbating the unequal distribution of access to health-care technologies within and among wealthy and low-income countries [2]. Furthermore, the digital divide could exacerbate inequitable access to health-care technologies by geography, gender, age or availability of devices, if countries do not take appropriate measures. Inappropriate use of AI could also perpetuate or exacerbate bias. Use of limited, low-quality, non-representative data in AI could perpetuate and deepen prejudices and disparities in health care. Biased inferences, misleading data analyses and poorly designed health applications and tools could be harmful. Predictive algorithms based on inadequate or inappropriate data can result in significant racial or ethnic bias. Use of high-quality, comprehensive datasets is essential.

AI could present a singular opportunity to augment and improve the capabilities of over-stretched health-care workers and providers. Yet, the introduction of AI for health care, as in many other sectors of the global economy, could have a significant negative impact on the health-care workforce. It could reduce the size of the workforce, limit, challenge or degrade the skills of health workers, and oblige them to retrain to adapt to the use of AI. Centuries of medical practice are based on relationships between provider and patient, and particular care must be taken when introducing AI technologies so that they do not disrupt such relationships.

The Universal Declaration of Human Rights, which includes pillars of patient rights such as dignity, privacy, confidentiality and informed consent, might be dramatically redefined or undermined as digital technologies take hold and expand. The performance of AI depends (among other factors) on the nature, type and volume of data and associated information and the conditions under which such data were gathered. The pursuit of data, whether by government or companies, could undermine privacy and autonomy at the service of government or private surveillance or commercial profit.

If privacy and autonomy are not assured, the resulting limitation of the ability to exercise the full range of human rights, including civil and political rights (such as freedom of movement and expression) and social and economic rights (such as access to health care and education), might have a wider impact.

AI technologies, like many information technologies used in health care, are usually designed by companies or through public-private partnerships (PPPs), although many governments also develop and deploy these technologies. Some of the world's largest technology companies are developing new applications and services, which they either own or invest in. Many of these companies have already accumulated large quantities of data, including health data, and exercise significant power in society and the economy. While these companies may offer innovative approaches, there is concern that they might eventually exercise too much power in relation to governments, providers and patients.

AI technologies are also changing where people access health care. AI technologies for health are increasingly distributed outside regulated health-care settings, including at the workplace, on social media and in the education system. With the rapid proliferation and evolving uses of AI for health care, including in response to the COVID-19 pandemic, government agencies, academic institutions, foundations, nongovernmental organizations and national ethics committees are defining how governments and other entities should use and regulate such technologies effectively. Ethically optimized tools and applications could sustain widespread use of AI to improve human health and the quality of life, while mitigating or eliminating many risks and bad practices.

To date, there is no comprehensive international guidance on use of AI for health in accordance with ethical norms and human rights standards. Most countries do not have laws or regulations to regulate use of AI technologies for health care, and their existing laws may not be adequate or specific enough for this purpose. WHO recognizes that ethics guidance based on the shared perspectives of the different entities that develop, use or oversee such technologies is critical to build trust in these technologies, to guard against negative or erosive effects and to avoid the proliferation of contradictory guidelines. Harmonized ethics guidance is therefore essential for the design and implementation of AI for global health.

The primary readership of this guidance document is ministries of health, as it is they that determine how to introduce, integrate and harness these technologies for the public good while restricting or prohibiting inappropriate use. The development, adoption and use of AI nevertheless requires an integrated, coordinated approach among government ministries beyond that for health. The stakeholders also include regulatory agencies, which must validate and define whether, when and how such technologies are to be used, ministries of education that teach current and future health-care workforces how such technologies function and are to be integrated into everyday practice, ministries of information technology that should facilitate the appropriate collection and use of health data and narrow the digital divide and countries' legal systems that should ensure that people harmed by AI technologies can seek redress.

This guidance document is also intended for the stakeholders throughout the health-care system who will have to adapt to and adopt these technologies, including medical researchers, scientists, health-care workers and, especially, patients. Access to such technologies can empower people who fall ill but can also leave them vulnerable, with fewer services and less protection. People have always been at the centre at all levels of decision-making in health care, whereas the inevitable growth of AI for health care could eventually challenge human primacy over medicine and health.

This guidance is also designed for those responsible for the design, deployment and refinement of AI technologies, including technologists and software developers. Finally, it is intended to guide the companies, universities, medical associations and international organizations that will, with governments and ministries of health, set policies and practices to define use of AI in the health sector. In identifying the many ethical concerns raised by AI and by providing the relevant ethical frameworks to address such concerns, this document is intended to support responsible use of AI worldwide.

AI is a fast-moving, evolving field and that many applications, not yet envisaged, will emerge as ever-greater public and private investment is dedicated to the use of AI for health. For example, in 2020, WHO issued interim guidance on the use of proximity tracking applications intended to facilitate contact-tracing during the COVID-19 pandemic. WHO may consider specific guidance for additional tools and applications and periodically update this guidance to keep pace with this rapidly changing field.

2 Artificial intelligence

"Artificial intelligence" generally refers to the performance by computer programs of tasks that are commonly associated with intelligent beings. The basis of AI is algorithms, which are translated into computer code that carries instructions for rapid analysis and transformation of data into conclusions, information or other outputs. Enormous quantities of data and the capacity to analyse such data rapidly fuel AI [3]. A specific definition of AI in a recommendation of the Council on Artificial Intelligence of the OECD [4] states:

> An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The various types of AI technology include machine-learning applications such as pattern recognition, natural language processing, signal processing and expert systems. Machine learning, which is a subset of AI techniques, is based on use of statistical and mathematical modelling techniques to define and analyse data. Such learned patterns are then applied to perform or guide certain tasks and make predictions.

Machine learning can be subcategorized according to how it learns from data into supervised learning, unsupervised learning and reinforced learning. In supervised learning, data used to train the model are labelled (the outcome variable is known), and the model infers a function from the data that can be used for predicting outputs from different inputs. Unsupervised learning does not involve labelling data but involves identification of hidden patterns in the data by a machine. Reinforcement learning involves machine learning by trial and error to achieve an objective for which the machine is "rewarded" or "penalized", depending on whether its inferences reach or hinder achievement of an objective [5]. Deep learning, also known as "deep structured learning", is a family of machine learning based on use of multi-layered models to progressively extract features from data. Deep learning can be supervised, unsupervised or semi-supervised. Deep learning generally requires large amounts of data to be fed into the model.

Many machine-learning approaches are data-driven. They depend on large amounts of accurate data, referred to as "big data", to produce tangible results. "Big data" are complex data that are rapidly collected in such unprecedented quantities that terabytes (one trillion units [bytes] of digital information), petabytes (1000 terabytes) or even zettabytes (one million petabytes) of storage space may be required as well as unconventional methods for their handling. The unique properties of big data are defined by four dimensions: volume, velocity, veracity and variety.

AI could improve the delivery of health care, such as prevention, diagnosis and treatment of disease [6], and is already changing how health services are delivered in several high-income countries (HIC). The possible applications of AI for health and medicine are expanding continually, although the use of AI may be limited outside HIC because of inadequate infrastructure. The applications can be defined according to the specific goals of use of AI and how AI is used to achieve those goals (methods). In health care, usable data have proliferated as a result of collection from numerous sources, including wearable technologies, genetic information generated by genome sequencing, electronic health-care records, radiological images and even from hospital rooms [7].

3 Applications of artificial intelligence for health

This section identifies AI technologies developed and used in HIC, although examples of such technologies are emerging (and being pilot-tested or used) in LMIC. Digital health technologies are already used widely in LMIC, including for data collection, dissemination of health information by mobile phones and extended use of electronic medical records on open-software platforms and cloud computing [8]. Schwabe and Wahl [9] have identified four uses of AI for health in LMIC: diagnosis, morbidity or mortality risk assessment, disease outbreak and surveillance, and health policy and planning.

3.1 In health care

The use of AI in medicine raises notions of AI replacing clinicians and human decision-making. The prevailing sentiment is, however, that AI is increasingly improving diagnosis and clinical care, based on earlier definitions of the role of computers in medicine [10] and regulations in which AI is defined as a support tool (to improve judgement).

3.1.1 Diagnosis and prediction-based diagnosis

AI is being considered to support diagnosis in several ways, including in radiology and medical imaging. Such applications, while more widely used than other AI applications, are still relatively novel, and AI is not yet used routinely in clinical decision-making. Currently, AI is being evaluated for use in radiological diagnosis in oncology (thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging and dose optimization for radiological treatment), in non-radiological applications (dermatology, pathology), in diagnosis of diabetic retinopathy, in ophthalmology and for RNA and DNA sequencing to guide immunotherapy [11]. In LMIC, AI may be used to improve detection of tuberculosis in a support system for interpreting staining images [12] or for scanning X-rays for signs of tuberculosis, COVID-19 or 27 other conditions [13].

Nevertheless, few such systems have been evaluated in prospective clinical trials. A recent comparison of deep-learning algorithms with health-care professionals in detection of diseases by medical imaging showed that AI is equivalent to human medical judgement in specific domains and applications in specific contexts but also that "few studies present externally validated results or compare the performance of deep learning models and health-care professionals using the same sample" [14]. Other questions are whether the performance of AI can be generalized to implementation in practice and whether AI trained for use in one context can be used accurately and safely in a different geographical region or context.

As AI improves, it could allow medical providers to make faster, more accurate diagnoses. AI could be used for prompt detection of conditions such as stroke, pneumonia, breast cancer by imaging [15, 16], coronary heart disease by echocardiography [17] and detection of cervical cancer [18]. Unitaid, a United Nations agency for improving diagnosis and treatment of infectious diseases in LMIC, launched a partnership with the Clinton Health Access Initiative in 2018 to pilot-test use of an AI-based tool to screen for cervical cancer in India, Kenya, Malawi, Rwanda, South Africa and Zambia [19]. Many low-income settings facing chronic shortages of health-care workers require assistance in diagnosis and assessment and to reduce their workload. It has been suggested that AI could fill gaps in the absence of health-care services or skilled workers [9].

AI might be used to predict illness or major health events before they occur. For example, an AI technology could be adapted to assess the relative risk of disease, which could be used for prevention of lifestyle diseases such as cardiovascular disease ([20], [21]) and diabetes [22]. Another use of AI for prediction could be to identify individuals with tuberculosis in LMIC who are not reached by the health system and therefore do not know their status [23]. Predictive analytics could avert other causes of unnecessary morbidity and mortality in LMIC, such as birth asphyxia. An expert system used in LMIC is 77% sensitive and 95% specific for predicting the need for resuscitation [8]. Several ethical challenges to prediction-based health care are discussed in section 6.5.

3.1.2 Clinical care

Clinicians might use AI to integrate patient records during consultations, identify patients at risk and vulnerable groups, as an aid in difficult treatment decisions and to catch clinical errors. In LMIC, for example, AI could be used in the management of antiretroviral therapy by predicting resistance to HIV drugs and disease progression, to help physicians optimize therapy [23]. Yet, clinical experience and knowledge about patients is essential, and AI will not be a substitute for clinical due diligence for the foreseeable future. If it did, clinicians might engage in "automation bias" and not consider whether an AI technology meets their needs or those of the patient. (See section 6.4.)

The wider use of AI in medicine also has technological challenges. Although many prototypes developed in both the public and the private sectors have performed well in field tests, they often cannot be translated, commercialized or deployed. An additional obstacle is constant changes in computing and information technology management, whereby systems become obsolete ("software erosion") and companies disappear. In resource-poor countries, the lack of digital infrastructure and the digital divide (See section 6.2.) will limit use of such technologies.

Health-care workers will have to adapt their clinical practice significantly as use of AI increases. AI could automate tasks, giving doctors time to listen to patients, address their fears and concerns and ask about unrelated social factors, although they may still worry about their responsibility and accountability. Doctors will have to update their competence to communicate risks, make predictions and discuss trade-offs with patients and also express their ethical and legal concern about understanding AI technology. Even if technology makes the predicted gains, those gains will materialize only if the individuals who manage health systems use them to extend the capacity of the health system in other areas, such as better availability of medicines or other prescribed interventions or forms of clinical care.

3.1.3 Emerging trends in the use of AI in clinical care

Several important changes imposed by the use of AI in clinical care extend beyond the provider-patient relationship. Four trends described here are: the evolving role of the patient in clinical care; the shift from hospital to home-based care; use of AI to provide "clinical" care outside the formal health system; and use of AI for resource allocation and prioritization. Each of these trends has ethical implications, as discussed below.

The evolving role of the patient in clinical care

AI could eventually change how patients self-manage their own medical conditions, especially chronic diseases such as cardiovascular diseases, diabetes and mental problems [24]. Patients already take significant responsibility for their own care, including taking medicines, improving their nutrition and diet, engaging in physical activity, caring for wounds or delivering injections. AI could assist in self-care, including through conversation agents (e.g., "chat bots"), health monitoring and risk prediction tools and technologies designed specifically for individuals with disabilities [24]. While a shift to patient-based care may be considered empowering and beneficial for some patients, others might find the additional responsibility stressful, and it might limit an individual's access to formal health-care services.

The growing use of digital self-management applications and technologies also raises wider questions about whether such technologies should be regulated as clinical applications, thus requiring greater regulatory scrutiny, or as "wellness applications", requiring less regulatory scrutiny. Many digital self-management technologies arguably fall into a "grey zone" between these two categories and may present a risk if they are used by patients for their own disease management or clinical care but remain largely unregulated or could be used without prior medical advice. Such concerns are exacerbated by the distribution of such applications by entities that are not a part of the formal health-care system. This related but separate trend is discussed below.

The shift from hospital to home-based care

Telemedicine is part of a larger shift from hospital- to home-based care, with use of AI technologies to facilitate the shift. They include remote monitoring systems, such as video-observed therapy for tuberculosis and virtual assistants to support patient care. Even before the COVID-19 pandemic, over 50 health-care systems in the USA were making use of telemedicine services [25]. COVID-19, having discouraged people in many settings from visiting health-care facilities, accelerated and expanded the use of telemedicine in 2020, and the trend is expected to continue. In China, the number of telemedicine providers has increased by nearly four times during the pandemic [26].

The shift to home-based care has also partly been facilitated by increased use of search engines (which rely on algorithms) for medical information as well as by the growth in the number of text or speech chatbots for health care [27], the performance of which has improved with improvements in natural language processing, a form of AI that enables machines to understand human language. The use of chatbots has also accelerated during the COVID-19 pandemic [28].

Furthermore, AI technologies may play a more active role in the management of patients' health outside clinical settings, such as in "just-in-time adaptive interventions". These rely on sensors to provide patients with specific interventions according to data collected previously and currently; they also notify a health-care provider of any emerging concern [29]. The growth and use of sensors and wearables may improve the effectiveness of "just-in-time adaptive interventions" but also raise concern, in view of the amount of data such technologies are collecting, how they are used and the burden such technologies may shift to patients.

Use of AI to extend "clinical" care beyond the formal health-care system

AI applications in health are no longer exclusively used in health-care systems (or home care), as AI technologies for health can be readily acquired and used by non-health system entities. This has meant that people can now obtain health-care services outside the health-care system. For example, AI applications for mental health are often provided through the education system, workplaces and social media and may even be linked to financial services [30]. While there may be support for such extended uses of health applications to compensate for both increased demand and a limited number of providers [31], they generate new questions and concerns. (See section 9.3.)

These three trends may require near-continuous monitoring (and self-monitoring) of people, even when they are not sick (or are "patients"). AI-guided technologies require the use of mobile health applications and wearables, and their use has increased with the trend to self-management [31]. Wearable technologies include those placed in the body (artificial limbs, smart implants), on the body (insulin pump patches, electroencephalogram devices) or near the body (activity trackers, smart watches and smart glasses). By 2025, 1.5 billion wearable units may be purchased annually.[1] Wearables will create more opportunities to monitor a person's health and to capture more data to predict health risks, often with greater efficiency and in a timelier manner.

Although such monitoring of "healthy" individuals could generate data to predict or detect health risks or improve a person's treatment when necessary, it raises concern, as it permits near-constant surveillance and collection of excessive data that otherwise should remain unknown or uncollected. Such data collection also contributes to the ever-growing practice of "biosurveillance", a form of surveillance for health data and other biometrics, such as facial features, fingerprints, temperature and pulse [32]. The growth of biosurveillance poses significant ethical and legal concerns, including the use of such data for medical and non-medical purposes for which explicit consent might not have been obtained or the repurposing of such data for non-health purposes by a government or company, such as within criminal justice or immigration systems. (See section 6.3.) Thus, such data should be liable to the same levels of data protection and security as for data collected on an individual in a formal clinical care setting.

---

[1] Presentation by Christian Stammel. Wearable Technologies, Germany, to the WHO Meeting of the Expert Group on Ethics and Governance of AI for Health, 6 March 2020.

Use of AI for resource allocation and prioritization

AI is being considered for use to assist in decision-making about prioritization or allocation of scarce resources. Prognostic scoring systems have long been available in critical care units. One of the best-known, Sequential Organ Failure Assessment (SOFA) [33], for analysis of the severity of illness and for predicting mortality, has been in use for decades, and SOFA scores have been widely used in some jurisdictions to guide allocation of resources for COVID-19 [34]. It is not an AI system; however, an AI version, "DeepSOFA" [35], has been developed.

The growing attraction of this use of AI has been due partly to the COVID-19 pandemic, as many institutions lack bed capacity and others have inadequate ventilators. Thus, hospitals and clinics in the worst-affected countries have been overwhelmed. It has been suggested that machine-learning algorithms could be trained and used to assist in decisions to ration supplies, identify which individuals should receive critical care or when to discontinue certain interventions, especially ventilator support [36]. AI tools could also be used to guide allocation of other scarce health resources during the COVID-19 pandemic, such as newly approved vaccines for which there is an insufficient initial supply [37].

Several ethical challenges associated with the use of AI for resource allocation and prioritization are described in section 6.5.

3.2 In health research and drug development

3.2.1 Application of AI for health research

An important area of health research with AI is based on use of data generated for electronic health records. Such data may be difficult to use if the underlying information technology system and database do not discourage the proliferation of heterogeneous or low-quality data. AI can nevertheless be applied to electronic health records for biomedical research, quality improvement and optimization of clinical care. From electronic health records, AI that is accurately designed and trained with appropriate data can help to identify clinical best practices before the customary pathway of scientific publication, guideline development and clinical support tools. AI can also assist in analysing clinical practice patterns derived from electronic health records to develop new clinical practice models.

A second (of many) application of AI for health research is in the field of genomics. Genomics is the study of the entire genetic material of an organism, which in humans consists of an estimated three billion DNA base pairs. Genomic medicine is an emerging discipline based on individuals' genomic information to guide clinical care and personalized approaches to diagnosis and treatment [38]. As the analysis of such large datasets is complex, AI is expected to play an important role in genomics. In health research, for example, AI could improve human understanding of disease or identify new disease biomarkers [38], although the quality of the data and whether they are representative and unbiased (See section 6.6.) could undermine the results.

3.2.2 Uses of AI in drug development

AI is expected in time to be used to both simplify and accelerate drug development. AI could change drug discovery from a labour-intensive to a capital- and data-intensive process with the use of robotics and models of genetic targets, drugs, organs, diseases and their progression, pharmacokinetics, safety and efficacy. AI could be used in drug discovery and throughout drug development to shorten the process and make it less expensive and more effective [39]. AI was used to identify potential treatments for Ebola virus disease, although, as in all drug development, identification of a lead compound may not result in safe, effective therapy [40].

In December 2020, DeepMind announced that its AlphaFold system had solved what is known as the "protein folding problem", in that the system can reliably predict the three-dimensional shape of a protein [41]. Although this achievement is only one part of a long process in understanding diseases and developing new medicines and vaccines, it should help to speed the development of new

medicines and improve the repurposing of existing medicines for use against new viruses and new diseases [41]. While this advance could significantly accelerate drug discovery, there is ethical concern about ownership and control of an AI technology that could be critical to drug development, as it might eventually be available to government, not-for-profit, academic and LMIC researchers only under commercial terms and conditions that limit its diffusion and use.

At present, drug development is led either by humans or by AI with human oversight. In the next two decades, as work with machines is optimized, the role of AI could evolve. Computing is starting to facilitate drug discovery and development by finding novel leads and evaluating whether they meet the criteria for new drugs, structuring unorganized data from medical imaging, searching large volumes of data, including health-care records, genetics data, laboratory tests, the Internet of Things, published literature and other types of health big data to identify structures and features, while recreating the body and its organs on chips (tissue chips) for AI analysis ([39], [42]). By 2040, testing of medicines might be virtual – without animals or humans – based on computer models of the human body, tumours, safety, efficacy, epigenetics and other parameters. Prescription drugs could be designed for each person. Such efforts could contribute to precision medicine or health care that is individually tailored to a person's genes, lifestyle and environment.

3.3 In health systems management and planning

Health systems, even in a single-payer, government-run system, may be overly complex and involve numerous actors who contribute to, pay for or benefit from the provision of health-care services. The management and administration of care may be laborious. AI can be used to assist personnel in complex logistical tasks, such as optimization of the medical supply chain, to assume mundane, repetitive tasks or to support complex decision-making. Some possible functions of AI for health systems management include: identifying and eliminating fraud or waste, scheduling patients, predicting which patients are unlikely to attend a scheduled appointment and assisting in identification of staffing requirements [43].

AI could also be useful in complex decision-making and planning, including in LMIC. For example, researchers in South Africa applied machine-learning models to administrative data to predict the length of stay of health workers in underserved communities [9]. In a study in Brazil, researchers used several government data sets and AI to optimize the allocation of health-system resources by geographical location according to current health challenges [9]. Allocation of scarce health resources through use of AI has raised concern, however, that resources may not be fairly allocated due, for example, to bias in the data. (See section 6.5.)

3.4 In public health and public health surveillance

Several AI tools for population and public health can be used in public health programmes. For example, new developments in AI could, after rigorous evaluation, improve identification of disease outbreaks and support surveillance. Several concerns about the use of technology for public health surveillance, promotion and outbreak response must, however, be considered before use of AI for such purposes, including the tension between the public health benefits of surveillance and ethical and legal concern about individual (or community) privacy and autonomy [44].

3.4.1 Health promotion

AI can be used for health promotion or to identify target populations or locations with "high-risk" behaviour and populations that would benefit from health communication and messaging (micro-targeting). AI programmes can use different forms of data to identify such populations, with varying accuracy, to improve message targeting.

Micro-targeting can also, however, raise concern, such as that with respect to commercial and political advertising, including the opaqueness of processes that facilitate micro-targeting. Furthermore, users who receive such messages may have no explanation or indication of why they

have been targeted [45]. Micro-targeting also undermines a population's equal access to information, can affect public debate and can facilitate exclusion or discrimination if it is used improperly by the public or private sector.

3.4.2 Disease prevention

AI has also been used to address the underlying causes of poor health outcomes, such as risks related to environmental or occupational health. AI tools can be used to identify bacterial contamination in water treatment plants, simplify detection and lower the costs. Sensors can also be used to improve environmental health, such as by analysing air pollution patterns or using machine learning to make inferences between the physical environment and healthy behaviour [29]. One concern with such use of AI is whether it is provided equitably or if such technologies are used only on behalf of wealthier populations and regions that have the relevant infrastructure for its use [46].

3.4.3 Surveillance (including prediction-based surveillance) and emergency preparedness

AI has been used in public health surveillance for collecting evidence and using it to create mathematical models to make decisions. Technology is changing the types of data collected for public health surveillance by the addition of digital "traces", which are data that are not generated specifically for public health purposes (such as from blogs, videos, official reports and Internet searches). Videos (e.g., YouTube) are another "rich" source of information for health insights [47].

Characterization of digital traces as "health data" raises questions about the types of privacy protection or other safeguards that should be attached to such datasets if they are not publicly available. For example, the use of digital traces as health data could violate the data protection principle of "purpose limitation", that individuals who generate such data should know what their data will be used for at the point of collection [48].

Such use also raises questions of accuracy. Models are useful only when appropriate data are used. Machine-learning algorithms could be more valuable when augmented by digital traces of human activity, yet such digital traces could also negatively impact an algorithm's performance. Google Flu Trends, for example, was based on search engine queries about complications, remedies, symptoms and antiviral medications for influenza, which are used to estimate and predict influenza activity. While Google Flu Trends first provided relatively accurate predictions before those of the US Centers for Disease Control and Prevention, it overestimated the prevalence of flu between 2011 and 2013 because the system was not re-trained as human search behaviour evolved [49].

Although many public health institutions are not yet making full use of these sources of data, surveillance itself is changing, especially real-time surveillance. For example, researchers could detect a surge in cases of severe pulmonary disease associated with the use of electronic cigarettes by mining disparate online sources of information and using Health Map, an online data-mining tool [50]. Similarly, Microsoft researchers have found early evidence of adverse drug reactions from web logs with an AI system. In 2013, the company's researchers detected side-effects of several prescription drugs before they were found by the US Food and Drug Administration's warning system [51]. In 2020, the US Food and Drug Administration sponsored a "challenge", soliciting public submissions to develop computation algorithms for automatic detection of adverse events from publicly available data [52]. Despite its potential benefits, real-time data collection, like the collection and use of digital traces, could violate data protection rules if surveillance was not the purpose of its initial collection, which is especially likely when data collection is automated.

Before the COVID-19 pandemic, WHO had started to develop EPI-BRAIN, a global platform that will allow experts in data and public health to analyse large datasets for emergency preparedness and response. (See also section 7.1.) AI has been used to assist in both detection and prediction during the COVID-19 pandemic, although some consider that the techniques and programming developed will "pay dividends" only during a subsequent pandemic [49]. HealthMap first issued a short bulletin about a new type of pneumonia in Wuhan, China, at the end of December 2019 [49]. Since then, AI has been used to "now-cast" (assess the current state of) the COVID-19 pandemic [49], while, in some

countries, real-time data on the movement and location of people has been used to build AI models to forecast regional transmission dynamics and guide border checks and surveillance [53]. In order to determine how such applications should be used, an assessment should be conducted of whether they are accurate, effective and useful.

3.4.4 Outbreak response

The possible uses of AI for different aspects of outbreak response have also expanded during the COVID-19 pandemic. They include studying SARS-CoV2 transmission, facilitating detection, developing possible vaccines and treatments and understanding the socio-economic impacts of the pandemic [54]. Such use of AI was already tested during the pandemic of Ebola virus disease in West Africa in 2014, although the assumptions underlying use of AI technologies to predict the spread of the Ebola virus were based on erroneous views of how the virus was spreading ([55], [56]). While many possible uses of AI have been identified and used during the COVID-19 pandemic, their actual impact is likely to have been modest; in some cases, early AI screening tools for SARS-CoV2 "were utter junk" with which companies "were trying to capitalise on the panic and anxiety" [57].

New applications [58] are intended to support the off-line response, although not all may involve use of AI. These have included proximity tracking applications intended to notify users (and possibly health authorities) that they have been in the proximity (for some duration) of an individual who subsequently tested positive for SARS-CoV2. Concern has been raised about privacy and the utility and accuracy of proximity-tracking applications, and WHO issued interim guidance on the ethical use of proximity-tracking applications in 2020 [59].

WHO and many ministries of health have also deployed symptom checkers, which are intended to guide users through a series of questions to assist in determining whether they should seek additional medical advice or testing for SARS-CoV2. The first symptom checkers were "hard coded", based on accumulated clinical judgement, as there were no previous data, and on a simple decision tree from older AI techniques, which involved direct encoding of expert knowledge. AI systems based on machine learning require accurate training, while data are initially scarce for a new disease such as COVID-19 [60]. New symptom checkers are based on machine learning to provide advice to patients [61], although their effectiveness is not yet known; all symptom checkers require that users provide accurate information.

AI has also been introduced to map the movements of individuals in order to approximate the effectiveness of government-mandated orders to remain in confinement, and, in some countries, AI technology has been used to identify individuals who should self-quarantine and be tested. These technologies raise legal and ethical concerns about privacy and risk of discrimination and also about possibly unnecessary restriction of movement or access to services, which heavily impact the exercise of a range of human rights [53]. As for all AI technologies, their actual effectiveness depends on whether the datasets are representative of the populations in which the technologies are used, and they remain questionable without systematic testing and evaluation. The uses described above are therefore not yet established.

3.5 The future of artificial intelligence for health

While AI may not replace clinical decision-making, it could improve decisions made by clinicians. In settings with limited resources, AI could be used to conduct screening and evaluation if insufficient medical expertise is available, a common challenge in many resource-poor settings. Yet, whether AI can advance beyond narrow tasks depends on numerous factors beyond the state of AI science and on the trust of providers, patients and health-care professionals in AI-based technologies. In the following sections of this document, ethical concerns and risks associated with the expanding use of AI for health are discussed, including by whom and how such technologies are deployed and developed. Technological, legal, security and ethical challenges and concerns are discussed not to dissuade potential use of AI for health but to ensure that AI fulfils its great potential and promise.

4 Laws, policies and principles that apply to artificial intelligence for health

Laws, policies and principles for regulating and managing the use of AI and specifically use of AI for health are fragmented and limited. Numerous principles and guidelines have been developed for application of "ethical" AI in the private and public sectors and in research institutions [62]; however, there is no consensus on its definition, best practices or ethical requirements, and different legal regimes and governance models are associated with each set of principles. Other norms, rules and frameworks also apply to use of AI, including human rights obligations, bioethics laws and policies, data protection laws and regulatory standards. These are summarized below and discussed elsewhere in the document. Section 5 provides a set of guiding principles agreed by the WHO Expert Group by consensus, on which this analysis and findings are based.

4.1 Artificial intelligence and human rights

Efforts to enumerate human rights and to fortify their observance through explicit legal mechanisms are reflected in international and regional human rights conventions, including the Universal Declaration on Human Rights, the International Covenant on Economic, Social and Cultural Rights (including General Comment No. 14, which defines the right to health), the International Covenant on Civil and Political Rights and regional human rights conventions, such as the African Charter on Human and People's Rights, the American Convention on Human Rights and the European Convention on Human Rights. Not all governments have acceded to key human rights instruments; some have signed but not ratified such charters or have expressed reservations to certain provisions. In general, however, human rights listed in international instruments establish a baseline for the protection and promotion of human dignity worldwide and are enforced through national legislation such as constitutions or human rights legislation.

Machine-learning systems could advance human rights but could also undermine core human rights standards. The Office of the High Commissioner for Human Rights has issued several opinions on the relation of AI to the realization of human rights. In guidance issued in March 2020, the Office noted that AI and big data can improve the human right to health when "new technologies are designed in an accountable manner" and could ensure that certain vulnerable populations have efficient, individualized care, such as assistive devices, built-in environmental applications and robotics [63]. The Office also noted, however, that such technologies could dehumanize care, undermine the autonomy and independence of older persons and pose significant risks to patient privacy – all of which are contrary to the right to health [63].

In February 2021, in a speech to the Human Rights Council, the United Nations Secretary-General noted a number of concerns for human rights associated with the growing collection and use of data on the COVID-19 pandemic and called on governments to "place human rights at the centre of regulatory frameworks and legislation on the development and use of digital technologies" [64]. Human rights organizations have interpreted and, when necessary, adapted existing human rights laws and standards to AI assessment and are reviewing them in the face of the challenges and opportunities associated with AI. The Toronto Declaration [65] addresses the impact of AI on human rights and situates AI within the universally binding, actionable framework of human rights laws and standards; it provides mechanisms for public and private sector accountability and the protection of people from discrimination and promotes equity, diversity and inclusion, while safeguarding equality and effective redress and remedy.

In 2018, the Council of Europe's Committee of Ministers issued draft recommendations to Member States on the impact of algorithmic systems on human rights [66]. The Council of Europe is further examining the feasibility and potential elements of a legal framework for the development, design and application of digital technologies according to its standards on human rights, democracy and the rule of law.

Legal frameworks for human rights, bioethics and privacy adopted by countries are applicable to several aspects of AI for health. They include Article 8 of the European Convention on Human Rights: the right to respect for private and family life, home and correspondence [67]; the Oviedo Convention on Human Rights and Biomedicine, which covers ethical principles of individual human rights and responsibilities [68]; the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data [69] and guidelines on the protection of individuals with regard to the processing of personal data in a world of big data, prepared by the Consultative Committee of Convention 108+ [69].

Yet, even with robust human rights standards, organizations and institutions recognize that better definition is required of how human rights standards and safeguards relate and apply to AI and that new laws and jurisprudence are required to address the interaction of AI and human rights. New legal guidance has been prepared by the Council of Europe. In 2019-2020, the Council established the Ad-hoc Committee on Artificial Intelligence to conduct broad multi-stakeholder consultations in order to determine the feasibility and potential elements of a legal framework for the design and application of AI according to the Council of Europe's standards on human rights, democracy and the rule of law. Further, in 2019, the Council of Europe released Guidelines on artificial intelligence and data protection [70], also based on the protection of human dignity and safeguarding human rights and fundamental freedom. In addition, the ethical charter of the European Commission for Efficiency of Justice includes five principles relevant to use of AI for health [71].

4.2 Data protection laws and policies

Data protection laws are "rights-based approaches" that provide standards for regulating data processing that both protect the rights of individuals and establish obligations for data controllers and processors. Data protection laws also increasingly recognize that people have the right not to be subject to decisions guided solely by automated processes. Over 100 countries have enacted data protection laws. One well-known set of data protection laws is the General Data Protection Regulation (GDPR) of the European Union (EU); in the USA, the Health Insurance Portability and Accountability Act, enacted in 1996, applies to privacy and to the security of health data.

Some standards and guidelines are designed specifically to manage the use of personal data for AI. For example, the Ibero-American Data Protection Network, which consists of 22 data protection authorities in Portugal and Spain and in Mexico and other countries in Central and South America and the Caribbean, has issued General Recommendations for the Processing of Personal Data in Artificial Intelligence [72] and specific guidelines for compliance with the principles and rights that govern the protection of personal data in AI projects [73].

4.3 Existing laws and policies related to health data

Several types of laws and policies govern the collection, processing, analysis, transfer and use of health data. The Council of Europe's Committee of Ministers issued a recommendation to Member States on the protection of health-related data in 2019 [74], and the African Union's convention on cybersecurity and personal data protection [2014] [75] requires that personal data involving genetic information and health research be processed only with the authorization of the national data protection authority through the Personal Data Protection Guidelines for Africa [76]. Generally, the African continent's digital transformation strategy [77] encourages African Union Member States to "have adequate regulation; particularly around data governance and digital platforms, to ensure that trust is preserved in the digitalization". In February 2021, the African Academy of Sciences and the African Union Development Agency released recommendations for data and biospecimen governance in Africa to promote a participant-centred approach to research involving human participants, while enabling ethical research practices on the continent and providing guidelines for governance [78].

Laws that govern the transfer of data among countries include those defined in trade agreements, intellectual property (IP) rules for the ownership of data and the role of competition law and policy related to the accumulation and control of data (including health data). These are discussed in detail later in this document.

4.4 General principles for the development and use of artificial intelligence

An estimated 100 proposals for AI principles have been published in the past decade, and studies have been conducted to identify which principles are most cited [79]. In one study of mapping and analysis of current principles and guidelines for ethical use of AI, convergence was found on transparency, justice, fairness, non-maleficence and responsibility, while other principles such as privacy, solidarity, human dignity and sustainability were under-represented [62].

Several intergovernmental organizations and countries have proposed such principles (Box 1).

Box 1 – Examples of AI ethics principles proposed by intergovernmental organizations and countries

The Recommendations of the OECD Council on Artificial Intelligence [80], the first intergovernmental standard on AI, were adopted in May 2019 by OECD's 36 member countries and have since been applied by a number of partner economies. The OECD AI principles [81] provided the basis for the AI principles endorsed by G20 governments in June 2019 [82]. While OECD recommendations are not legally binding, they carry a political commitment and have proved highly influential in setting international standards in other policy areas (e.g., privacy and data protection) and helping governments to design national legislation. The OECD launched an online platform for public policy on AI, the AI Policy Observatory [83] (See section 9.6.) and is cooperating on this and other initiatives on the ethical implications of AI with the Council of Europe, the United Nations Economic, Scientific and Cultural Organization (UNESCO) and WHO.

– In 2019, the Council of Europe Commissioner for Human Rights issued recommendations to ensure that human rights are strengthened rather than undermined by AI: Unboxing artificial intelligence: 10 steps to protect human rights recommendations [84].
– The European Commission appointed 52 representatives from academia, civil society and industry to its High-level Expert Group on Artificial Intelligence and issued Ethics Guidelines for Trustworthy AI [85].
– Japan has issued several guidelines on the use of AI, including on research and development and utilization [86].
– China has issued National Governance Principles for the New Generation Artificial Intelligence, which serves as the national principles for AI governance in China [87]. Academia and industry have jointly issued the Beijing Artificial Intelligence Principles [88].2
– In Singapore, a series of initiatives on AI governance and ethics was designed to build an ecosystem of trust to support adoption of AI. They include Asia's first Model AI governance framework, released in January 2019; an international industry-led Advisory Council on the Ethical Use of AI and Data formed in June 2018; a research programme on the governance of AI and data use established in partnership with the Singapore Management University in September 2018 [89]; and a certification programme for ethics and governance of AI for companies and developers [90].
– The African Union's High-level Panel on Emerging Technologies is preparing broad guidance on the use of AI to promote economic development and its use in various sectors, including health care [91].

2 Presentation by Professor Yi Zeng, Chinese Academy of Sciences, 4 October 2019, to the WHO working group on ethics and governance of AI for health.

4.5 Principles for use of artificial intelligence for health

No specific ethical principles for use of AI for health have yet been proposed for adoption worldwide. Before WHO's work on guidance on the ethics and governance of AI for health, the WHO Global Conference on Primary Health Care issued the Astana Declaration [92], which includes principles for the use of digital technology. The Declaration calls for promotion of rational, safe use and protection of personal data and use of technology to improve access to health care, enrich health service delivery, improve the quality of service and patient safety and increase the efficiency and coordination of care.

UNESCO has guidance and principles for the use of AI in general and for the use of big data in health. UNESCO's work on the ethical implications of AI is supported by two standing expert committees, the World Commission on the Ethics of Scientific Knowledge and Technology and the International Bioethics Committee. Other work includes the report of the International Bioethics Committee on big data and health in 2017, which identified important elements of a governance framework [93]; the World Commission on the Ethics of Scientific Knowledge and Technology report on robotics ethics in 2017 [94]; a preliminary study on the ethics of AI by UNESCO in 2019, which raised ethical concern about education, science and gender [95]; a recommendation on the ethics of AI to be considered by UNESCO's General Conference in 2021; and a report by the World Commission on the Ethics of Scientific Knowledge and Technology on the Internet of Things.

In 2019, the United Kingdom's National Health Service (NHS) released a code of conduct, with 10 principles for the development and use of safe, ethical, effective, data-based health and care technologies [96]. In October 2019, The Lancet and The Financial Times launched a joint commission, The Governing Health Futures 2030: Growing up in a Digital World Commission, on the convergence of digital health, AI and universal health coverage, which will consult between October 2019 and December 2021 [97].

4.6 Bioethics laws and policies

Bioethics laws and policies play a role in regulating the use of AI, and several bioethics laws have been revised in recent years to include recognition of the growing use of AI in science, health care and medicine. The French Government's most recent revision of its national bioethics law [98], which was endorsed in 2019, establishes standards to address the rapid growth of digital technologies in the health-care system. It includes standards for human supervision, or human warranty, that require evaluation by patients and clinicians at critical points in the development and deployment of AI. It also supports free, informed consent for the use of data and the creation of a secure national platform for the collection and processing of health data.

4.7 Regulatory considerations

Regulation of AI technologies is likely to be developed and implemented by health regulatory authorities responsible for ensuring the safety, efficacy and appropriate use of technologies for health care and therapeutic development. A WHO expert group that is preparing considerations for the regulation of AI for health has discussed areas that should be considered by stakeholders, including developers and regulators, in examining new AI technologies. They include documentation and transparency, risk management and the life-cycle approach, data quality, analytical and clinical validation, engagement and collaboration, and privacy and data protection. Many regulatory authorities are preparing considerations and frameworks for the use of AI, and they should be examined, potentially with the relevant regulatory agency. Governance of AI through regulatory frameworks and the ethical principles that should be considered are discussed in section 9.5.

5 Key ethical principles for use of artificial intelligence for health

Ethical principles for the application of AI for health and other domains are intended to guide developers, users and regulators in improving and overseeing the design and use of such technologies. Human dignity and the inherent worth of humans are the central values upon which all other ethical principles rest.

An ethical principle is a statement of a duty or a responsibility in the context of the development, deployment and continuing assessment of AI technologies for health. The ethical principles described below are grounded in basic ethical requirements that apply to all persons and that are considered noncontroversial. The requirements are as follows.

– Avoid harming others (sometimes called "Do no harm" or nonmaleficence).

– Promote the well-being of others when possible (sometimes called "beneficence"). Risks of harm should be minimized, while maximizing benefits. Expected risks should be balanced against expected benefits.

– Ensure that all persons are treated fairly, which includes the requirement to ensure that no person or group is subject to discrimination, neglect, manipulation, domination or abuse (sometimes called "justice" or "fairness").

– Deal with persons in ways that respect their interests in making decisions about their lives and their person, including health-care decisions, according to informed understanding of the nature of the choice to be made, its significance, the person's interests and the likely consequences of the alternatives (sometimes called "respect for persons" or "autonomy").

Additional moral requirements can be derived from this list of fundamental moral requirements. For example, safeguarding and protecting individual privacy is not only recognized as a legal requirement in many countries but is also important to enable people to control sensitive information about themselves and self-determination (respect for their autonomy) and to avoid harm.

These ethical principles are intended to provide guidance to stakeholders about how basic moral requirements should direct or constrain their decisions and actions in the specific context of developing, deploying and assessing the performance of AI technologies for health. These principles are also intended to emphasize issues that arise from the use of a technology that could alter relations of moral significance. For example, it has long been recognized that health-care providers have a special duty to advance these values with respect to patients because of the centrality of health to individual well-being, because of the dependence of patients on health professionals for information about their diagnosis, prognosis and the relative merits of the available treatment or prevention options, and the importance of free and open exchange of information to the provider-patient relationship. If AI systems are used by health-care workers to conduct clinical tasks or to delegate clinical tasks that were once reserved for humans, programmers who design and program such AI technologies should also adhere to these ethical obligations.

Thus, the ethical principles are important for all stakeholders who seek guidance in the responsible development, deployment and evaluation of AI technologies for health, including clinicians, systems developers, health system administrators, policy-makers in health authorities, and local and national governments. The ethical principles listed here should encourage and assist governments and public sector agencies to keep pace with the rapid evolution of AI technologies through legislation and regulation and should empower medical professionals to use AI technologies appropriately.

Ethical principles should also be embedded within professional and technological standards for AI. Software engineers already are guided by standards such as for fitness for purpose, documentation and provenance, and version control. Standards are required to guide the interoperability and design of a program, for continuing education of those who develop and use such technologies and for governance. Moreover, the standards for the evaluation and external audit of systems are evolving in the context of their use. In health computing, there are standards for system integration, electronic health records, system interoperability, implementation and programming structures.

Although ethical principles do not always clearly address limitations in the uses of such technologies, governments should ban or restrict the use of AI or other technologies if they violate or imperil the exercise of human rights, do not conform to other principles or regulations or would be introduced in unprepared or other inappropriate contexts. For example, many countries lack data protection laws or have inadequate regulatory frameworks to guide the introduction of AI technologies.

The claim that certain basic moral requirements must constrain and guide the conduct of persons can also be expressed in the language of human rights. Human rights are intended to capture a basic set of moral and legal requirements for conduct to which every person is entitled regardless of race, sex, nationality, ethnicity, language, religion or any other feature. These rights include human dignity, equality, non-discrimination, privacy, freedom, participation, solidarity and accountability.

Machine-learning systems could advance the protection and enforcement of human rights (including the human right to health) but could undermine core human rights such as non-discrimination and privacy. Human rights and ethical principles are intimately interlinked; because human rights are legally binding, they provide a powerful framework by which governments, international organizations and private actors are obligated to abide. Private sector actors have the responsibility to respect human rights, independently of state obligations. In fulfilling this responsibility, private sector actors must take continuous proactive and reactive steps to ensure that they do not abuse or contribute to the abuse of human rights.

The existence of a human rights framework does not, however, obviate the need for continuing ethical deliberation. Indeed, much of ethics is intended to expand upon and complement the norms and obligations established in human rights agreements. In many situations, multiple ethical considerations are relevant and require weighing up and balancing to accommodate the multiple principles at stake. An ethically acceptable decision depends on consideration of the full range of appropriate ethical considerations, ensuring that multiple perspectives are factored into the analysis and creating a decision-making process that stakeholders will consider fair and legitimate.

This guidance identifies six ethical principles to guide the development and use of AI technology for health. While ethical principles are universal, their implementation may differ according to the cultural, religious and other social context. Many of the ethical issues arising in the use of AI and machine learning are not completely new but have arisen for other applications of information and communication technologies for health, such as use of any computer to track a disease or make a diagnosis or prognosis. Computers were performing these tasks with various programs long before AI became noteworthy. Ethical guidance and related principles have been articulated for fields such as telemedicine and data-sharing. Likewise, several ethical frameworks have been developed for AI in general, outside the health sector. (See section 4.) The ethical principles listed here are those identified by the WHO Expert Group as the most appropriate for the use of AI for health.

5.1 Protect autonomy

Adoption of AI can lead to situations in which decision-making could be or is in fact transferred to machines. The principle of autonomy requires that any extension of machine autonomy not undermine human autonomy.[3] In the context of health care, this means that humans should remain in full control of health-care systems and medical decisions. AI systems should be designed demonstrably and systematically to conform to the principles and human rights with which they cohere; more specifically, they should be designed to assist humans, whether they be medical providers or patients, in making informed decisions. Human oversight may depend on the risks associated with an AI system but should always be meaningful and should thus include effective,

[3] Building on the work of W.D. Ross (99), Beauchamp and Childress (100) formulated a principle-based approach to bioethics in which they added a "principle of respect for autonomy" to Ross' three other principles. The Principles of Biomedical Ethics (100), although highly influential, is not universally accepted as dispositive.

transparent monitoring of human values and moral considerations. In practice, this could include deciding whether to use an AI system for a particular health-care decision, to vary the level of human discretion and decision-making and to develop AI technologies that can rank decisions when appropriate (as opposed to a single decision). These practices can ensure a clinician can override decisions made by AI systems and that machine autonomy can be restricted and made "intrinsically reversible".

Respect for autonomy also entails the related duties to protect privacy and confidentiality and to ensure informed, valid consent by adopting appropriate legal frameworks for data protection. These should be fully supported and enforced by governments and respected by companies and their system designers, programmers, database creators and others. AI technologies should not be used for experimentation or manipulation of humans in a health-care system without valid informed consent. The use of machine-learning algorithms in diagnosis, prognosis and treatment plans should be incorporated into the process for informed and valid consent. Informed and valid consent means that essential services are not circumscribed or denied if an individual withholds consent and that additional incentives or inducements should not be offered by either a government or private parties to individuals who do provide consent.

Data protection laws are one means of safeguarding individual rights and place obligations on data controllers and data processors. Such laws are necessary to protect privacy and the confidentiality of patient data and to establish patients' control over their data. Construed broadly, data protection laws should also make it easy for people to access their own health data and to move or share those data as they like. Because machine learning requires large amounts of data – big data – these laws are increasingly important.

5.2 Promote human well-being, human safety and the public interest

AI technologies should not harm people. They should satisfy regulatory requirements for safety, accuracy and efficacy before deployment, and measures should be in place to ensure quality control and quality improvement. Thus, funders, developers and users have a continuous duty to measure and monitor the performance of AI algorithms to ensure that AI technologies work as designed and to assess whether they have any detrimental impact on individual patients or groups.

Preventing harm requires that use of AI technologies does not result in any mental or physical harm. AI technologies that provide a diagnosis or warning that an individual cannot address because of lack of appropriate, accessible or affordable health care should be carefully managed and balanced against any "duty to warn" that might arise from incidental and other findings, and appropriate safeguards should be in place to protect individuals from stigmatization or discrimination due to their health status.

5.3 Ensure transparency, explainability and intelligibility

AI should be intelligible or understandable to developers, users and regulators. Two broad approaches to ensuring intelligibility are improving the transparency and explainability of AI technology.

Transparency requires that sufficient information (described below) be published or documented before the design and deployment of an AI technology. Such information should facilitate meaningful public consultation and debate on how the AI technology is designed and how it should be used. Such information should continue to be published and documented regularly and in a timely manner after an AI technology is approved for use.

Transparency will improve system quality and protect patient and public health safety. For instance, system evaluators require transparency in order to identify errors, and government regulators rely on transparency to conduct proper, effective oversight. It must be possible to audit an AI technology, including if something goes wrong. Transparency should include accurate information about the assumptions and limitations of the technology, operating protocols, the properties of the data

(including methods of data collection, processing and labelling) and development of the algorithmic model.

AI technologies should be explainable to the extent possible and according to the capacity of those to whom the explanation is directed. Data protection laws already create specific obligations of explainability for automated decision-making. Those who might request or require an explanation should be well informed, and the educational information must be tailored to each population, including, for example, marginalized populations. Many AI technologies are complex, and the complexity might frustrate both the explainer and the person receiving the explanation. There is a possible trade-off between full explainability of an algorithm (at the cost of accuracy) and improved accuracy (at the cost of explainability).

All algorithms should be tested rigorously in the settings in which the technology will be used in order to ensure that it meets standards of safety and efficacy. The examination and validation should include the assumptions, operational protocols, data properties and output decisions of the AI technology. Tests and evaluations should be regular, transparent and of sufficient breadth to cover differences in the performance of the algorithm according to race, ethnicity, gender, age and other relevant human characteristics. There should be robust, independent oversight of such tests and evaluation to ensure that they are conducted safely and effectively.

Health-care institutions, health systems and public health agencies should regularly publish information about how decisions have been made for adoption of an AI technology and how the technology will be evaluated periodically, its uses, its known limitations and the role of decision-making, which can facilitate external auditing and oversight.

5.4 Foster responsibility and accountability

Humans require clear, transparent specification of the tasks that systems can perform and the conditions under which they can achieve the desired level of performance; this helps to ensure that health-care providers can use an AI technology responsibly. Although AI technologies perform specific tasks, it is the responsibility of human stakeholders to ensure that they can perform those tasks and that they are used under appropriate conditions.

Responsibility can be assured by application of "human warranty", which implies evaluation by patients and clinicians in the development and deployment of AI technologies. In human warranty, regulatory principles are applied upstream and downstream of the algorithm by establishing points of human supervision. The critical points of supervision are identified by discussions among professionals, patients and designers. The goal is to ensure that the algorithm remains on a machine-learning development path that is medically effective, can be interrogated and is ethically responsible; it involves active partnership with patients and the public, such as meaningful public consultation and debate [101]. Ultimately, such work should be validated by regulatory agencies or other supervisory authorities.

When something does go wrong in application of an AI technology, there should be accountability. Appropriate mechanisms should be adopted to ensure questioning by and redress for individuals and groups adversely affected by algorithmically informed decisions. This should include access to prompt, effective remedies and redress from governments and companies that deploy AI technologies for health care. Redress should include compensation, rehabilitation, restitution, sanctions where necessary and a guarantee of non-repetition.

The use of AI technologies in medicine requires attribution of responsibility within complex systems in which responsibility is distributed among numerous agents. When medical decisions by AI technologies harm individuals, responsibility and accountability processes should clearly identify the relative roles of manufacturers and clinical users in the harm. This is an evolving challenge and remains unsettled in the laws of most countries. Institutions have not only legal liability but also a

duty to assume responsibility for decisions made by the algorithms they use, even if it is not feasible to explain in detail how the algorithms produce their results.

To avoid diffusion of responsibility, in which "everybody's problem becomes nobody's responsibility", a faultless responsibility model ("collective responsibility"), in which all the agents involved in the development and deployment of an AI technology are held responsible, can encourage all actors to act with integrity and minimize harm. In such a model, the actual intentions of each agent (or actor) or their ability to control an outcome are not considered.

5.5 Ensure inclusiveness and equity

Inclusiveness requires that AI used in health care is designed to encourage the widest possible appropriate, equitable use and access, irrespective of age, gender, income, ability or other characteristics. Institutions (e.g., companies, regulatory agencies, health systems) should hire employees from diverse backgrounds, cultures and disciplines to develop, monitor and deploy AI. AI technologies should be designed by and evaluated with the active participation of those who are required to use the system or will be affected by it, including providers and patients, and such participants should be sufficiently diverse. Participation can also be improved by adopting open-source software or making source codes publicly available.

AI technology – like any other technology – should be shared as widely as possible. AI technologies should be available not only in HIC and for use in contexts and for needs that apply to high-income settings but they should also be adaptable to the types of devices, telecommunications infrastructure and data transfer capacity in LMIC. AI developers and vendors should also consider the diversity of languages, ability and forms of communication around the world to avoid barriers to use. Industry and governments should strive to ensure that the "digital divide" within and between countries is not widened and ensure equitable access to novel AI technologies.

AI technologies should not be biased. Bias is a threat to inclusiveness and equity because it represents a departure, often arbitrary, from equal treatment. For example, a system designed to diagnose cancerous skin lesions that is trained with data on one skin colour may not generate accurate results for patients with a different skin colour, increasing the risk to their health.

Unintended biases that may emerge with AI should be avoided or identified and mitigated. AI developers should be aware of the possible biases in their design, implementation and use and the potential harm that biases can cause to individuals and society. These parties also have a duty to address potential bias and avoid introducing or exacerbating health-care disparities, including when testing or deploying new AI technologies in vulnerable populations.

AI developers should ensure that AI data, and especially training data, do not include sampling bias and are therefore accurate, complete and diverse. If a particular racial or ethnic minority (or other group) is underrepresented in a dataset, oversampling of that group relative to its population size may be necessary to ensure that an AI technology achieves the same quality of results in that population as in better-represented groups.

AI technologies should minimize inevitable power disparities between providers and patients or between companies that create and deploy AI technologies and those that use or rely on them. Public sector agencies should have control over the data collected by private health-care providers, and their shared responsibilities should be defined and respected. Everyone – patients, health-care providers and health-care systems – should be able to benefit from an AI technology and not just the technology providers. AI technologies should be accompanied by means to provide patients with knowledge and skills to better understand their health status and to communicate effectively with health-care providers. Future health literacy should include an element of information technology literacy.

The effects of use of AI technologies must be monitored and evaluated, including disproportionate effects on specific groups of people when they mirror or exacerbate existing forms of bias and

discrimination. Special provision should be made to protect the rights and welfare of vulnerable persons, with mechanisms for redress if such bias and discrimination emerges or is alleged.

5.6 Promote artificial intelligence that is responsive and sustainable

Responsiveness requires that designers, developers and users continuously, systematically and transparently examine an AI technology to determine whether it is responding adequately, appropriately and according to communicated expectations and requirements in the context in which it is used. Thus, identification of a health need requires that institutions and governments respond to that need and its context with appropriate technologies with the aim of achieving the public interest in health protection and promotion. When an AI technology is ineffective or engenders dissatisfaction, the duty to be responsive requires an institutional process to resolve the problem, which may include terminating use of the technology.

Responsiveness also requires that AI technologies be consistent with wider efforts to promote health systems and environmental and workplace sustainability. AI technologies should be introduced only if they can be fully integrated and sustained in the health-care system. Too often, especially in under-resourced health systems, new technologies are not used or are not repaired or updated, thereby wasting scare resources that could have been invested in proven interventions. Furthermore, AI systems should be designed to minimize their ecological footprints and increase energy efficiency, so that use of AI is consistent with society's efforts to reduce the impact of human beings on the earth's environment, ecosystems and climate. Sustainability also requires governments and companies to address anticipated disruptions to the workplace, including training of health-care workers to adapt to use of AI and potential job losses due to the use of automated systems for routine health-care functions and administrative tasks.

6 Ethical challenges to use of artificial intelligence for health care

Several ethical challenges are emerging with the use of AI for health, many of which are especially relevant to LMIC. These challenges must be addressed if AI technologies are to support achievement of universal health coverage. Use of AI to extend health-care coverage and services in marginalized communities in HIC can raise similar ethical concerns, including an enduring digital divide, lack of good-quality data, collection of data that incorporate clinical biases (as well as inappropriate data collection practices) and lack of treatment options after diagnosis.

6.1 Assessing whether artificial intelligence should be used

There are risks of overstatement of what AI can accomplish, unrealistic estimates of what could be achieved as AI evolves and uptake of unproven products and services that have not been subjected to rigorous evaluation for safety and efficacy [93]. This is due partly to the enduring appeal of "technological solutionism", in which technologies such as AI are used as a "magic bullet" to remove deeper social, structural, economic and institutional barriers [102]. The appeal of technological solutions and the promise of technology can lead to overestimation of the benefits and dismissal of the challenges and problems that new technologies such as AI may introduce. This can result in an unbalanced health-care policy and misguided investments by countries that have few resources and by HIC that are under pressure to reduce public expenditure on health care [103]. It can also divert attention and resources from proven but underfunded interventions that would reduce morbidity and mortality in LMIC.

First, the AI technology itself may not meet the standards of scientific validity and accuracy that are currently applied to medical technologies. For example, digital technologies developed in the early stages of the COVID-19 pandemic did not necessarily meet any objective standard of efficacy to justify their use [104]. AI technologies have been introduced as part of the pandemic response without adequate evidence, such as from randomized clinical trials, or safeguards [9]. An emergency does not justify deployment of unproven technologies [104]; in fact, efforts to ensure that resources were

allocated where they were most urgently needed should have heightened the vigilance of both companies and governments (such as regulators and ministries of health) to ensure that the technologies were accurate and effective.

Secondly, the benefits of AI may be overestimated when erroneous or overly optimistic assumptions are made about the infrastructure and institutional context in which the technologies will be used and where the intrinsic requirements for use of the technology cannot be met. In some low-income countries, financial resources and information and communication technology infrastructure lag those of HIC, and the significant investments that would be required might discourage use. This is discussed in greater detail in section 6.2. The quality and availability of data may not be adequate for use of AI, especially in LMIC. There is a danger that poor-quality data will be collected for AI training, which may result in models that predict artefacts in the data instead of actual clinical outcomes. There may also be no data, which, with poor-quality data, could distort the performance of an algorithm, resulting in inaccurate performance, or an AI technology might not be available for a specific population because of insufficient usable data. Additionally, significant investment may be required to make non-uniform data sets collected in LMIC usable. Compilation of data in resource-poor settings is difficult and time-consuming, and the additional burden on community health workers should be considered. Data are unlikely to be available on the most vulnerable or marginalized populations, including those for whom health-care services are lacking, or they might be inaccurate. Data may also be difficult to collect because of language barriers, and mistrust may lead people to provide incorrect or incomplete information. Often, irrelevant data are collected, which can undermine the overall quality of a dataset.[4] Broader concern about the collection and use of data, as well as bias in data, is discussed below.

There may not be appropriate or enforceable regulations, stakeholder participation or oversight, all of which are required to ensure that ethical and legal concerns can be addressed and human rights are not violated. For example, AI technologies may be introduced in countries without up-to-date data protection and confidentiality laws (especially for health-related data) or without the oversight of data protection authorities to rigorously protect confidentiality and the privacy of individuals and communities. Furthermore, regulatory agencies in LMIC may not have the capacity or expertise to assess AI technologies to ensure that systematic errors do not affect diagnosis, surveillance and treatment.

Thirdly, there may be enough ethical concern about a use case or a specific AI technology, even if it provides accurate, useful information and insights, to discourage a particular use. An AI technology that can predict which individuals are likely to develop type 2 diabetes or HIV infection could provide benefits to an at-risk individual or community but could also give rise to unnecessary stigmatization of individuals or communities, whose choices and behaviour are questioned or even criminalized, result in over-medicalization of otherwise healthy individuals, create unnecessary stress and anxiety and expose individuals to aggressive marketing by pharmaceutical companies and other for-profit health-care services [105]. Furthermore, certain AI technologies, if not deployed carefully, could exacerbate disparities in health care, including those related to ethnicity, socioeconomic status or gender.

Fourthly, like all new heath technologies, even if an AI technology does not trigger an ethics warning, its benefits may not be justified by the extra expense or cost (beyond information and communication technology infrastructure) associated with the procurement, training and technology investment required [43]. Robotic surgery may produce better outcomes, but the opportunity costs associated with the investment must also be considered.

---

[4] Presentation by Dr Amel Ghoulia, Bill & Melinda Gates Foundation, 3 October 2019, to the WHO working group on ethics and governance of AI for health.

Fifthly, enough consideration may not be given to whether an AI technology is appropriate and adapted to the context of LMIC, such as diverse languages and scripts in a country or among countries [9]. Lack of investment in, for example, translation can mean that certain applications do not operate correctly or simply cannot be used by a population. Such lack of foresight points to a wider problem, which is that many AI technologies are designed by and for high-income populations and by individuals or companies with inadequate understanding of the characteristics of the target populations in LMIC.

Unrealistic expectations of what AI can achieve may, however, unnecessarily discourage its use. Thus, machines and algorithms (and the data used for algorithms) are expected in the public imagination to be perfect, while humans can make mistakes. Medical professionals might overestimate their ability to perform tasks and ignore or underestimate the value of algorithmic decision tools, for which the challenges can be managed and for which evidence indicates a measurable benefit. Not using the technology could result in avoidable morbidity and mortality, making it blameworthy not to use a certain AI technology, especially if the standard of care is already shifting to its use [106]. For medical professionals to make such an assessment, they require greater transparency with regard to the performance and utility of AI technologies, a principle enumerated in section 5 of this document, as well as effective regulatory oversight. The role of regulatory agencies in ensuring rigorous testing, transparent communication of outcomes and monitoring of performance is discussed in section 9.5.

Even after an AI technology has been introduced into a health-care system, its impact should be evaluated continuously during its real-world use, as should the performance of an algorithm if it learns from data that are different from its training data.

Impact assessments can also guide a decision on use of AI in an area of health before and after its introduction [106]. (See section 7.3.) Assessment of whether to introduce an AI technology in a low-income country or resource-poor setting may lead to a different conclusion from such an assessment in a high-income setting. Risk-benefit calculations that do not favour a specific use of AI in HIC may be interpreted differently for a low-income country that lacks, for example, enough health-care workers to perform certain tasks or which would otherwise forego use of more accurate diagnostic instruments, such that individuals receive inaccurate diagnoses and the wrong treatment.

The use of AI to resource-poor contexts should, however, be extended carefully to avoid situations in which large numbers of people receive accurate diagnoses of a health condition but have no access to appropriate treatment. Health-care workers have a duty to provide treatment after testing for and confirmation of disease, and the relatively low cost at which AI diagnostics can be deployed should be accompanied by careful planning to ensure that people are not left without treatment.[5] Prediction tools for anticipating a disease outbreak will have to be complemented by robust surveillance systems and other effective measures.

6.2 Artificial intelligence and the digital divide

Many LMIC have sophisticated economies and digital infrastructure, while others, such as India, have both world-class digital infrastructure and millions of people without electricity. The countries with the greatest challenges to adoption of AI are classified as least developed; however, AI could allow those countries to leapfrog existing models of health-care delivery to improve health outcomes [23].

One challenge that could affect the uptake of AI is the "digital divide", which refers to uneven distribution of access to, use of or effect of information and communication technologies among any number of distinct groups. Although the cost of digital technologies is falling, access has not become

[5] The International Council of Nurses noted: "Ethical issues may arise if there is the capability of AI diagnostics but not the capacity to provide treatment. Issues like this have arisen in the field of endoscopy in some countries where some diagnostic services for screening are withheld because of the limited access to surgical services." Communication from the International Council of Nurses to WHO on 6 January 2021.

more equitable. For example, 1.2 billion women (327 million fewer women than men) in LMIC do not use mobile Internet services because they cannot afford to or do not trust the technology, even though the cost of the devices should continue to fall [107]. Gender is only one dimension of the digital divide; others are geography, culture, religion, language and generation. The digital divide begets other disparities and challenges, many of which affect the use of AI, and AI itself can reinforce and exacerbate the disparity. Thus, in 2019, the United Nations Secretary-General's High-level Panel on Digital Cooperation [108] recommended that

> by 2030, every adult should have affordable access to digital networks, as well as digitally enabled financial and health services, as a means to make a substantial contribution to achieving the Sustainable Development Goals.

The human and technical resources required to realize the benefits of digital technologies fully are also unequally distributed, and infrastructure to operate digital technologies may be limited or inexistent. Some technologies require an electricity grid and information and communication technology infrastructure, including electrification, Internet connectivity, wireless and mobile networks and devices. Solar energy may provide a path forward for many countries if the climate is appropriate, as investment is increasing and the cost of solar energy has decreased dramatically in the past decade [109]. Nevertheless, at present, an estimated 860 million people worldwide do not have access to electricity, including 600 million people in sub-Saharan Africa, and there is growing pressure on the electrical grid in cities due to urbanization [110]. Even in high-income economies with near-universal electrification and enough resources, the digital divide has persisted. In the USA, for example, millions of people in rural areas and in cities still lack access to high-speed broadband services, and 60% of health-care facilities outside metropolitan areas also lack broadband [111].

Even as countries overcome the digital divide, technology providers should be required to provide infrastructure, services and programs that are interoperable, so that different platforms and applications can work seamlessly with one another, as well as affordable devices (for example, smartphones) that do not require consumers to trade privacy for affordability [112]. This will ensure that the emerging digital health-care system is not fragmented and is equitable.

6.3 Data collection and use

The collection, analysis and use of health data, including from clinical trials, laboratory results and medical records, is the bedrock of medical research and the practice of medicine. Over the past two decades, the data that qualify as health data have expanded dramatically. They now include massive quantities of personal data about individuals from many sources, including genomic data, radiological images, medical records and non-health data converted into health data [113]. The various types of data, collectively known as "biomedical big data", form a health data ecosystem that includes data from standard sources (e.g., health services, public health, research) and further sources (environmental, lifestyle, socioeconomic, behavioural and social). See Figure 1 [114].



Figure 1 – Health data ecosystem [115]

Thus, there are many more sources of health data, entities that wish to make use of such data and commercial and non-commercial applications. The development of a successful AI system for use in health care relies on high-quality data for both training the algorithm and validating the algorithmic model.

The potential benefits of biomedical big data can be ethically important, as AI technologies based on high-quality data can improve the speed and accuracy of diagnosis, improve the quality of care and reduce subjective decision-making. The ubiquity of health data and the potential sensitivity of health care to data indicate possible benefits. Health care is still lagging in the adoption of data science and AI as compared with other sectors (although some would disagree), and individuals informed of the potential benefits of the collection and use of such data might support use of such data for their personal benefit or that of a wider group.[6]

---

6 Presentation by Dr Andrew Morris, Health Data Research United Kingdom, 3 October 2019 to the WHO working group on ethics and governance of AI for health.

Several concerns may undermine effective use of health data in AI-guided research and drug development. Concern about the use of health data is not limited to their use in AI, although AI has exacerbated the problem. One concern with health data is their quality, especially with those from LMIC (see above). Furthermore, training data will always have one or more systemic biases because of under-representation of a gender, age, race, sexual orientation or other characteristic. These biases will emerge during modelling and subsequently diffuse through the resulting algorithm [103]. Concern about the impact of bias is discussed in section 6.6.

A second major concern is safeguarding individual privacy. The collection, use, analysis and sharing of health data have consistently raised broad concern about individual privacy, because lack of privacy may either harm an individual (such as future discrimination on the basis of one's health status) or cause a wrong, such as affecting a person's dignity if sensitive health data are shared or broadcast to others [116]. There is a risk that sharing or transferring data leaves them vulnerable to cyber-theft or accidental disclosure [116]. Recommendations generated by an algorithm from an individual's health data also raise privacy concerns, as a person may expect that such "new" health data are private [116], and it may be illegal for third parties to use "new" health data. Such privacy concerns are heightened for stigmatized and vulnerable populations, for whom data disclosure can lead to discrimination or punitive measures [117]. There is also concern about the rights of children [118], which could include future discrimination based on the data accumulated about a child, children's ability to protect their privacy and their autonomy to make choices about their health care. Measures to collect data or track an individual's status and to construct digital identities to store such information have accelerated during the COVID-19 pandemic. See Box 2.

Box 2 – The emergence of digital identification in the COVID-19 pandemic

The COVID-19 pandemic is expanding and accelerating the creation of infrastructure for digital identities to store health data for several uses. In China, a QR code system has been established from the digital payment system established by Alipay, a mobile and online payment platform, to introduce an "Alipay Health Code", in which the data collected are used to establish an algorithm to "draw automated conclusions as to whether someone is a contagion risk" [119]. For a national programme to vaccinate millions of people against SARS-Cov2, India may use its national digital ID system, Aadhar, to avoid duplication and to track beneficiaries [120]. Many entities around the world, including travel firms, airports, some governments and political leaders, as well as the digital ID industry, are calling for the introduction of immunity passports or a digital "credential given to a person who is assumed to be immune from SARS-CoV2 and so protected against re-infection" [121]. In some countries, technologies such as proximity-tracking applications have been credited with improving the response to the pandemic, because there was already a system in place to support the use of such technologies, effective communication, widespread adoption and a "social compact" between policy-makers and the public [122].

For many of these technologies, however, there is concern about whether they are effective (scientifically valid), whether they will create forms of discrimination or targeting of certain populations and whether they may exclude certain segments of the population or not be applicable by people who do not have access to the appropriate technology and infrastructure. They also raise concern about the generation of a permanent digital identity for individuals linked to their health and personal data, for which they may not have given consent, which could permanently undermine individual autonomy and privacy [123]. In particular, there is concern that governments could use such information to establish mass surveillance or scoring systems to monitor everyday activities, or companies could use such data and systems for other purposes [124].

A third major concern is that health data collected by technology providers may exceed what is required and that such excess data, so-called "behavioural data surplus" [125], is repurposed for uses that raise serious ethical, legal and human rights concerns. The uses might include sharing such data with government agencies so that they can exercise control or use punitive measures against individuals [104]. Such repurposing, or "function creep", is a challenge that predates but is heightened by the use of AI for health care. For example, in early 2021, the Singapore Government admitted that data obtained from its COVID-19 proximity-tracing application (Trace Together) could also be accessed "for the purpose of criminal investigation", despite prior assurances that this would not be

permitted [126]. In February 2021, legislation was introduced to restrict the use of such data for only the most "serious" criminal investigations, such as for murder or terrorism-related charges, with penalties for any unauthorized use [127].

Such data may also be shared with companies that use them to develop an AI technology for marketing goods and services or to create prediction-based products to be used, for example, by an insurance firm [128] or a large technology company. Such uses of health data, often unknown to those who have supplied the data, have generated front-page headlines and public concern [129]. The provision of health data to commercial entities has also resulted in the filing of legal actions by individuals whose health data (de-identified) have been disclosed on behalf of all affected individuals. See Box 3.

Box 3 – Dinerstein vs Google

Google announced a strategic partnership with the University of Chicago and the University of Chicago Medicine in the USA in May 2017 [130]. The aim of the partnership was to develop novel machine-learning tools to predict medical events such as unexpected hospital admissions. To realize this goal, the University shared hundreds of thousands of "de-identified" patients' records with Google. One of the University's patients, Matt Dinerstein, filed a class action complaint against the University and Google in June 2019 on behalf of all patients whose records were disclosed [131].

Dinerstein brought several claims, including breach of contract, against the University and Google, alleging prima facie violation of the US Health Insurance Portability and Accountability Act. According to an article published in 2018 by the defendants [132], the patients' medical records shared with Google "were de-identified, except that dates of service were maintained in the (…) dataset". The dataset also included "free-text medical notes" [132]. Dinerstein accused the defendants of insufficient anonymization of the records, putting the patients' privacy at risk. He alleged that the patients could easily be re-identified by Google by combining the records with other available data sets, such as geolocation data from Google Maps (by so-called "data triangulation"). Moreover, Dinerstein asserted that the University had not obtained express consent from each patient to share their medical records with Google, despite the technology giant's commercial interest in the data.

The issue of re-identification was largely avoided by the district judge, who dismissed Dinerstein's lawsuit in September 2020. The reasons given for dismissal included Dinerstein's failure to demonstrate damages that had occurred because of the partnership. This case illustrates the challenges of lawsuits related to data-sharing and highlights the lack of adequate protection of the privacy of health data. In the absence of ethical guidelines and adequate legislation, patients may have difficulty in maintaining control of their personal medical information, particularly in circumstances in which the data can be shared with third parties and in the absence of safeguards against re-identification.

This case study was written by Marcelo Corrales Compagnucci (CeBIL Copenhagen), Sara Gerke (Harvard Law School) and Timo Minssen (CeBIL Copenhagen).

Some companies have already collected large quantities of health data through their products and services, to which users voluntarily supply health data (user-generated health data) [133]. They may acquire further data through a data aggregator or broker [134] or may rely on governments to aggregate data that can be used by public, not-for-profit and private sector entities [135]. Such data may include "mundane" data that were not originally characterized as "health data"; however, machine learning can elicit sensitive details from such ordinary personal data and thus transform them into a special category of sensitive data [136] that may require protection.

Concern about the commercialization of health data includes individual loss of autonomy, a principle stated in section 5, loss of control over the data (with no explicit consent to such secondary use), how such data (or outcomes generated by such data) may be used by the company or a third party, with concern that companies are allowed to profit from the use of such data, and concern about privacy, as companies may not meet the duty of confidentiality, whether purposefully or inadvertently (for example due to a data breach) [137]. Thus, once an individual's medical history is exposed, it cannot be replaced in the same way as a new credit card can be obtained after a breach.

6.3.1 Data colonialism

A fourth concern with biomedical big data is that it may foster a divide between those who accumulate, acquire, analyse and control such data and those who provide the data but have little control over their use. This is especially true with respect to data collected from underrepresented groups, many of which are predominantly in LMIC, often with the broad ambition of collecting data for development or for humanitarian ends rather than to promote local economic development and governance [138]. Insufficient data from underrepresented groups affect them negatively, and attention has focused on either encouraging such groups to provide data or instituting measures to collect data. Generating more data from LMIC, however, also carries risks, including "data colonialism", in which the data are used for commercial or non-commercial purposes without due respect for consent, privacy or autonomy. Collection of data without the informed consent of individuals for the intended uses (commercial or otherwise) undermines the agency, dignity and human rights of those individuals; however, even informed consent may be insufficient to compensate for the power dissymmetry between the collectors of data and the individuals who are the sources. This is a particular concern because of the possibility that companies in countries with strict regulatory frameworks and data protection laws could extend data collection to LMIC without such control. While regulatory frameworks such as the EU's GDPR include an "extra-territorial" clause that requires compliance with its standards outside the EU, entities are not obliged to provide a right of redress as guaranteed under the EU GDPR, and companies may use such data but not provide appropriate products and services to the underserved communities and countries from which the data were obtained. Individuals in these regions therefore have little or no knowledge of how their data are being used, by a government or company, no opportunity to provide any form of consent for how the data could be used and often less bargaining power if recommendations based on the data have an adverse effect on an individual or a community [139].

6.3.2 Mechanisms for safeguarding privacy – Do they work?

When meaningful consent is possible, it can overcome many concerns, including those related to privacy. Yet, true informed consent is increasingly infeasible in an era of biomedical big data, especially in an environment driven mainly by companies seeking to generate profits from the use of data [113]. The scale and complexity of biomedical big data make it impossible to keep track of and make meaningful decisions about all uses of personal data [113]. All the potential uses of health data may not be known, as they may eventually be linked to and used for a purpose that is far removed from the original intention. Patients may be unable to consent to current and future uses of their health data, such as for population-level data analytics or predictive-risk modelling [113]. Even if a use lends itself to consent, the procedures may fall short, individuals might not be able to consent, such as because they have insufficient access to a health data system, or access to health care is perceived or actually denied if consent is not provided.

One concern is in the management of use of health data (probably collected for different purposes and not necessarily to support the use of AI) after an individual has died. Such data could provide numerous benefits for medical research [140], to improve understanding of the causes of cancer [141] or to increase the diversity of data used for medical AI. These data must, however, also be protected against unauthorized use. Existing laws either define limited circumstances in which such data can be used or restrict how they can be used [142]. In the GDPR, a data protection law does not apply to deceased persons, and, under Article 27, EU Member States "may provide for rules regarding the processing of personal data of deceased persons" [143]. Proposals have been made to improve the sharing of such data through voluntary and participatory approaches by which individuals can provide broad or selective consent for use of their data after death, much as individuals can provide consent for use of their organs for medical research [143].

If patients' privacy cannot be safeguarded by consent mechanisms, other privacy safeguards, including a data holder's duty of confidentiality, also have shortcomings. Although confidentiality is a well-recognized pillar of medical practice, the duty of confidentiality may not be sufficient to cover

the many types of data now used to guide AI health technologies and may also not be sufficient to control the production and transfer of health data [113].

A proactive approach to preserving privacy is de-identification or anonymization or pseudo-anonymization of health data. De-identification prevents connection of personal identifiers to information. Anonymization of personal data is a subcategory of de-identification whereby both direct and indirect personal identifiers are removed, and technical safeguards are used to ensure zero risk of re-identification, whereas de-identified data can be re-identified by use of a key [144]. Pseudo-anonymization is defined in Article 5 of the GDPR [145] as:

> processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

The use of such techniques could safeguard privacy and encourage data-sharing but also raises several concerns and challenges. In the USA for example, fully de-identified health data can be used for other purposes without consent [146]. De-identification may not always be successful, as "data triangulation" techniques can be used to reconstruct a de-identified, incomplete dataset by a third party for re-identification of an individual [147]. It may be impossible completely to de-identify some types of data, such as genome sequences, as relationships to other people whose identity and partial sequence are known can be inferred. Such relationships may allow direct identification of small groups and to narrow down identification to families ([128], [148]).

Anonymization may not be possible during health data collection. For example, in predictive AI, time-course data must be collected from a single individual at several times, obviating anonymization until data at all time points are collected. Furthermore, while anonymization may minimize the risks of (re-)identification of a person, it can reduce the positive benefits of health data, including re-assembly of fragments of an individual's health data into a comprehensive profile of a patient, which is required for some forms of AI such as predictive algorithms of mortality. Furthermore, anonymization may undermine a person's right to control their own data and how it may be used [113]. Other techniques could be used to preserve privacy, including differential privacy, synthetic data generation and k-anonymity, which are briefly discussed in section 7.1.

6.4 Accountability and responsibility for decision-making with artificial intelligence

This section addresses the challenges of assigning responsibility and accountability for the use of AI for health care, a guiding principle noted in section 5. Much of the momentum of AI is based on the notion that use of such technologies for diagnosis, care or systems could improve clinical and institutional decision-making for health care. Clinicians and health-care workers have numerous cognitive biases and commit diagnostic errors. The US National Academy of Sciences found that 5% of US adults who seek health advice receive erroneous diagnoses and that such errors account for 10% of all patient deaths [149]. At the institutional level, machine learning might reduce inefficiency and errors and ensure more appropriate allocation of resources, if the underlying data are both accurate and representative [149].

AI-guided decision-making also introduces several trade-offs and risks. One set of trade-offs is associated with the displacement of human judgement and control and concern about using AI to predict a person's health status or the evolution of disease. This is a major ethical and epistemological challenge to humans as the centre of production of knowledge and also to the system of production of knowledge for medicine. These considerations are addressed in section 6.5.

Governments can violate human rights (and companies can fail to respect human rights), undermine human dignity or cause tangible harm to human health and well-being by using AI-guided technologies. These violations may not be foreseen during development of an AI technology and may emerge only once the technology evolves in real-world use. If proactive measures such as greater

transparency and continuous updating of training data do not avoid harm, recourse may be made through civil (and occasionally criminal) liability. The use of liability regimes to address harm caused by AI-guided technologies is addressed in section 8.

Responsibility ensures that individuals and entities are held accountable for any adverse effects of their actions and is necessary to maintain trust and to protect human rights. Certain characteristics of AI technologies, however, affect notions of responsibility (and accountability), including their opacity, reliance on human input, interaction, discretion, scalability, capacity to generate hidden insights and the complexity of the software. One challenge to assigning responsibility is the 'control problem' associated with AI, wherein developers and designers of AI may not be held responsible, as AI-guided systems function independently of their developers and may evolve in ways that the developer could claim were not foreseeable [150]. This creates a responsibility gap, which could place an undue burden on a victim of harm or on the clinician or health-care worker who uses the technology but was not involved in its development or design ([150], [151]). Assigning responsibility to the developer might provide an incentive to take all possible steps to minimize harm to the patient. Such expectations are already well established for the producers of other commonly used medical technologies, including drug and vaccine manufacturers, medical device companies and medical equipment makers.

The 'control problem' will become ever more salient with the emergence of automated AI. Technology companies are making large investments in automating the programming of AI technologies, partly because of the scarcity of AI developers. Automation of AI programming, through programs such as BigML, Google AutoML and Data Robot, might be attractive to public health institutions that wish to use AI but lack the budget to hire AI developers [152]. While automated AI programming might be more accurate, its use might not be fair, ethical or safe in certain situations. If AI programming is automated, the checks and balances provided by the involvement of a human developer to ensure safety and identify errors would also be automated, and the control problem is abstracted one step further away from the patient.

A second challenge is the "many hands problem" or the "'traceability" of harm, which bedevils health-care decision-making systems [153] and other complex systems [154] even in the absence of AI. As the development of AI involves contributions from many agents, it is difficult, both legally and morally, to assign responsibility [150], which is diffused among all the contributors to the AI-guided technology. Participation of a machine in making decisions may also discourage assignment of responsibility to the humans involved in the design, selection and use of the technology [150]. Diffusion of responsibility may mean that an individual is not compensated for the harm he or she suffers, the harm itself and its cause are not fully detected, the harm is not addressed and societal trust in such technologies may be diminished if it appears that none of the developers or users of such technologies can be held responsible [155].

A third challenge to assigning responsibility is the issuance of ethics guidance by technology companies, separately or jointly [156]. Such guidance sets out norms and standards to which the companies commit themselves to comply publicly and voluntarily. Many companies have issued such guidance in the absence of authoritative or legally binding international standards. Recognition by technology companies that AI technologies for use in health care and other sectors are of public concern and must be carefully designed and deployed to avoid harm, such as violations of human rights or bodily injury, is welcome. Such guidelines may, however, depending on how they are implemented, be little more than "ethics washing" [150]. First, the public tends to have little or no role in setting such standards [157]. Secondly, such guidelines tend to apply to the prospective behaviour of companies for the technologies they design and deploy (role responsibility) and not historic responsibility for any harms for which responsibility should be allocated. This creates a responsibility gap, as it does not address causal responsibility or retrospective harm [150]. Thirdly, monitoring of whether companies are complying with their own guidance tends to be done internally, with little to no transparency, and without enforcement by institutions or mechanisms empowered to

act independently to evaluate whether the commitments are being met ([157], [158]). Finally, these commitments are not legally enforceable if violated [158].

AI provides great power and benefits (including the possibility of profit) to those who design and deploy such systems. Thus, reciprocity should apply – companies that reap direct and indirect benefits from AI-guided technologies should also have to shoulder responsibility for any negative consequences (section 8), especially as it is health-care providers who will bear the immediate brunt of any psychological stress if an AI technology causes harm to a patient. Companies should also allow independent audits and oversight of enforcement of its own ethics standards to ensure that the standards are being met and that corrective action is taken if a problem arises.

6.4.1 Accountability for AI-related errors and harm

Clinicians already use many non-AI technologies in diagnosis and treatment, such as X-rays and computer software. As AI technologies are used to assist or improve clinical decision-making and not to replace it, there may be an argument to initially hold clinicians accountable for any harm that results from their use in health care. In the same way as for non-AI technologies, however, this oversimplifies the reasons for harm and who should be held accountable for such harm. If a clinician makes a mistake in using the technology, he or she may be held accountable if they were trained in its use that otherwise may not have been included in their medical training [159]. Yet, if there is an error in the algorithm or the data used to train the AI technology, for example, accountability might be better placed with those who developed or tested the AI technology rather than requiring the clinician to judge whether the AI technology is providing useful guidance [159].

There are other reasons for not holding clinicians solely accountable for decisions made by AI technologies, several of which apply to assigning accountability for the use of non-AI health technologies. First, clinicians do not exercise control over an AI-guided technology or its recommendations [151]. Secondly, as AI technologies tend to be opaque and may use "black-box" algorithms, a physician may not understand how an AI system converts data into decisions [151]. Thirdly, the clinician may not have chosen to use the AI technology but does so because of the preferences of the hospital system or of other external decision-makers.

Furthermore, if physicians were made accountable for harm caused by an AI technology, technology companies and developers could avoid accountability, and human users of the technology would become the scapegoats of all faults arising from its use, with no control over the decisions made by the AI technology [150]. Furthermore, with the emergence of autonomous systems for driving and warfare, there is growing concern about whether humans can exert "meaningful control" over such technologies or whether the technologies will increasingly make decisions independently of human input. (See section 6.5.)

Clinicians should not, however, be fully exempt from accountability for errors in content, in order to avoid "automation bias" or lack of consideration of whether an automated technology meets their needs or those of the patient [159]. In automation bias, a clinician may overlook errors that should have been spotted by human-guided decision-making. While physicians must be able to trust an algorithm, they should not ignore their own expertise and judgement and simply rubber-stamp the recommendation of a machine [160]. Some AI technology may not issue a single decision but a set of options from which a physician must select. If the physician makes the wrong choice, what should the criteria be for holding the physician accountable?

Assignation of accountability is even more complex when a decision is made to use an AI technology throughout a health-care system, as the developer, the institution and the physician may all have played a role in the medical harm, yet none is fully to blame [149]. In such situations, accountability may rest not with the provider or the developer of the technology but with the government agency or institution that selected, validated and deployed it.

6.5 Autonomous decision-making

Decision-making has not yet been "fully transferred" from humans to machines in health care. While AI is used only to augment human decision-making in the practice of public health and medicine, epistemic authority has, in some circumstances, been displaced, whereby AI systems (such as with the use of computer simulations) are displacing humans from the centre of knowledge production ([161], [162]). Furthermore, there are signs of full delegation of routine medical functions to AI. Delegation of clinical judgement introduces concern about whether full delegation is legal, as laws increasingly recognize the right of individuals not to be subject to solely automated decisions when such decisions would have a significant effect. Full delegation also creates a risk of automation bias on the part of the provider, as discussed above. Other concerns could emerge if human judgement is increasingly replaced by machine-guided judgement, and wider ethical concern would arise with loss of human control, especially if prediction-based health care becomes the norm. Yet, as for autonomous cars, it is unlikely that AI in medicine will ever achieve full autonomy. It may achieve only conditional automation or require human back-up [163].

6.5.1 Implications of replacing human judgement for clinical care

There are benefits of replacing human judgement and of humans ceding control over certain aspects of clinical care. Humans could make worse decisions that are less fair and more biased compared to machines (concern about bias in the use of AI is discussed below). Use of AI systems to make specific, well-defined decisions may be entirely justified if there is compelling clinical evidence that the system performs the task better than a human. Leaving decisions to humans when machines can perform them more rapidly, more accurately and with greater sensitivity and specificity can mean that some patients suffer avoidable morbidity and mortality without the prospect of some offsetting benefit [106].

In some cases, automation of routine, mundane functions, such as recording information, could liberate a medical provider to build or enhance a relationship with a patient while AI-guided machines automate certain aspects of caregiving [24]. Other mundane functions could be fully assumed by AI, such as automatic adjustment of a hospital ward temperature.

The shift to applying AI technologies for more complex areas of clinical care will, however, present several challenges. One is the likely emergence of "peer disagreement" between two competent experts – an AI machine and a doctor [149]. In such situations, there is no means of combining the decisions or of reasoning with the algorithm, as it cannot be accessed or engaged to change its mind. There are also no clear rules for determining who is right, and if a patient is left to trust either a technology or a physician, the decision may depend on factors that have no basis in the "expertise" of the machine or the doctor. Choosing one of the two leads to an undesirable outcome. If the doctor ignores the machine, AI has added little value [149]. If the doctor accepts the machine's decision, it may undermine his or her authority and weaken their accountability. Some may argue that the recommendation of an algorithm should be preferred, as it combines the expertise of multiple experts and many data points [149].

The challenge of human-computer interactions has been addressed by validating systems, providing appropriate education for users and validating the systems continuously. It may, however, be ethically challenging for doctors to rely on the judgement of AI, as they have to accept decisions based on black-box algorithms [159]. The widely held convention is that many algorithms, e.g., those based on artificial neural networks or other complex models, are black boxes that make inferences and decisions that are not understood even by their developers [164]. It may therefore be questioned whether doctors can be asked to act on decisions made by such black-box algorithms.

AI should therefore be transparent and explainable, which is listed as a core guiding principle in section 5. Some argue that, if a trade-off must be made between even greater transparency (and explainability) and accuracy, transparency should be preferred. This requirement, however, goes beyond what may be possible or even desirable in a medical context. While it is often possible to

explain to a patient why a specific treatment is the best option for a specific condition, it is not always possible to explain how that treatment works or its mechanism of action, because some medical interventions are used before their mode of action is understood [165]. It may be more important to explain how a system has been validated and whether a particular use falls within the parameters with which the system can be expected to produce reliable results rather than explaining how an AI model arrives at a particular judgement [166]. Clinicians require other types of information, even if they do not understand exactly how an algorithm functions, including the data on which it was trained, how and who built the AI model and the variables underlying the AI model.

6.5.2 Implications of the loss of human control in clinical care

Loss of human control by assigning decision-making to AI-guided technologies could affect various aspects of clinical care and the health-care system. They include the patient, the clinician-patient relationship (and whether it interrupts communication between them), the relation of the health-care system to technology providers and the choices that societies should make about standards of care.

Although providing individuals with more opportunities to share data and to obtain autonomous health advice could improve their agency and self-care, it could also generate anxiety and fatigue [159]. As more personal data are collected by such technologies and used by clinicians, patients might increasingly be excluded from shared decision-making and left unable to exercise agency or autonomy in decisions about their health [149]. Most patients have insufficient knowledge about how and why AI technologies make certain decisions, and the technologies themselves may not be sufficiently transparent, even if a patient is well informed. In some situations, individuals may feel unable to refuse treatment, partly also because the patient cannot speak with or challenge the recommendation of an AI-guided technology (e.g., a notion that the "computer knows best") or is not given enough information or a rationale for providing informed consent [149].

Hospitals and health-care providers are unlikely to inform patients that AI was used as a part of decision-making to guide, validate or overrule a provider. There is, however, no precedent for seeking the consent of patients to use technologies for diagnosis or treatment. Nevertheless, the use of AI in medicine and failure to disclose its use could challenge the core of informed consent and wider public trust in health care. This challenge depends on whether any of the reasons for obtaining informed consent – protection, autonomy, prevention of abusive conduct, trust, self-ownership, non-domination and personal integrity – is triggered by the use of AI in clinical care [167]. See Box 4 for additional discussion on whether and how providers should disclose the use of AI for clinical care.

Box 4 – Informed consent during clinical care

Consider use of an AI in a hospital to make recommendations on a drug and dosage for a patient. The AI recommends a particular drug and dosage for patient A. The physician does not, however, understand how the AI reached its recommendation. The AI has a highly sophisticated algorithm and is thus a black box for the physician. Should the physician follow the AI's recommendation? If patients were to find out that an AI or machine-learning system was used to recommend their care but no one had told them, how would they feel? Does the physician have a moral or even a legal duty to tell patient A that he or she has consulted an AI technology? If so, what essential information should the physician provide to patient A? Should disclosure of the use of AI be part of obtaining informed consent and should a lack of sufficient information incur liability? [167]

Transparency is crucial to promoting trust among all stakeholders, particularly patients. Physicians should be frank with patients from the onset and inform them of the use of AI rather than hiding the technology. They should try their best to explain to their patients the purpose of using AI, how it functions and whether it is explainable. They should describe what data are collected, how they are used and shared with third parties and the safeguards for protection of patients' privacy. Physicians should also be transparent about any weaknesses of the AI technology, such as any biases, data breaches or privacy concerns. Only with transparency can the deployment of AI for health care and health science, including hospital practice and clinical trials [168], become a long-term success. Trust is key to facilitating the adoption of AI in medicine.

Note – This case study was written by Marcelo Corrales Compagnucci (CeBIL Copenhagen), Sara Gerke (Harvard Law School) and Timo Minssen (CeBIL Copenhagen).

Physicians who are left out of decision-making between a patient and an AI health technology may also feel loss of control, as they can no longer engage in the back-and-forth that is currently integral to clinical care and shared decision-making between providers and patients [160]. Some may consider loss of physician control over patients as promoting patient autonomy, but there is equally a risk of surrendering decision-making to an AI technology, which may be more likely if the technology is presented to the patient as providing better insight into their health status and prognosis than a physician [160].

Furthermore, if an AI technology reduces contact between a provider and a patient, it could reduce the opportunities for clinicians to offer health promotion interventions to the patient and undermine general supportive care, such as the benefits of human-human interaction when people are often at their most vulnerable [159]. Some AI technologies do not sever the relationship between doctor and patient but help to improve contact and communication, for example, by providing an analysis of different treatment options, which the doctor can talk through with the patient and explain the risks.

Loss of control could be construed as surrendering not just to a technology but also to companies that exert power over the development, deployment and use of AI for health care. At present, technology companies are investing resources to accumulate data, computing power and human resources to develop new AI health technologies ([169]-[171]). This may be done by large companies in partnership with the public sector, as in the United Kingdom [168], but could be done by concentrating different areas of expertise or decision-making in different companies, with the rules and standards of care governed by the companies that manage the technologies rather than health care systems. In China, several large technology companies, including Ping An [171], Tencent [174], Baidu [175] and Alibaba [176], are rapidly expanding the provision of both online and offline health services and new points of access to health care, backed by accumulation of data and use of AI. Companies, unlike health systems or governments, may, however, ignore the needs of citizens and the obligations owed to citizens, as there is a distinction between citizens and customers. These concerns heighten the importance of regulation and careful consideration of the role of companies in direct provision of health-care services.

6.5.3 The ethics of using AI for resource allocation and prioritization

Use of computerized decision-support programs – AI or not – to inform or guide resource allocation and prioritization for clinical care has long raised ethical issues [177]. They include managing conflicts between human and machine predictions, difficulty in assessing the quality and fitness for purpose of software, identifying appropriate users and the novel situation in which a decision for a patient is guided by a machine analysis of other patients' outcomes. In some situations, well-intentioned efforts to base decisions about allocations on an algorithm that relies only on a rules-based formula produce unintended outcomes. Such was the case in allocation of vaccines against COVID-19 at a medical institution in California, USA, on the basis of a rules-based formula in which very few of the available vaccine doses were allocated to those medical workers most at risk of contracting the virus, while prioritizing "higher-ranked" doctors at low-risk of COVID-19 [178].

Moreover, there is a familiar problem and risk that data in both traditional databases and machine-learning training sets might be biased. Such bias could lead to allocation of resources that discriminates against, for example, people of colour; decisions related to gender, ethnicity or socioeconomic status might similarly be biased. Such forms of bias and discrimination might not only be found in data but intentionally included in algorithms, such that formulas are written to discriminate against certain communities or individuals. At population level, this could encourage use of resources for people who will have the greatest net benefit, e.g., younger, healthier individuals, and divert resources and time from costly procedures intended for the elderly. Thus, if an AI technology is trained to "maximize global health", it may do so by allocating most resources to healthy people in order to keep them healthy and not to a disadvantaged population. This dovetails with a wider "conceptual revolution" in medicine, whereas

> twentieth-century medicine aimed to heal the sick. Twenty-first-century medicine is increasingly aimed to upgrade the healthy…. Consequently, by 2070 the poor could very well enjoy much better healthcare than today, but the gap separating them from the rich will nevertheless be much greater [179].

As more data are amassed and AI technologies are increasingly integrated into decision-making, providers and administrators will probably rely on the advice given (while guarding against automation bias). Yet, such technologies, if designed for efficiency of resource use, could compromise human dignity and equitable access to treatment. They could mean that decisions about whether to provide certain costly treatments or operations are based on predicted life span and on estimates of quality-adjusted life years or new metrics based on data that are inherently biased. In some countries in which AI is not used, patients are already triaged to optimize patient flow, and such decisions often affect those who are disadvantaged or powerless, such as the elderly, people of colour and those with genetic defects or disabilities.

Ethical design (see section 7.1) could mitigate these risks and ensure that AI technologies are used to assist humans by appropriate resource allocation and prioritization. Furthermore, such technologies must be maintained as a means of aiding human decision-making and assuring that humans ultimately make the right critical life-and-death decisions by adequately addressing the risks of such uses of AI and providing those affected by such decisions with contestation rights.

Use of AI tools for triage or rationing is one of the most compelling reasons for ensuring adequate governance or oversight. Although intentional harm is not ethically controversial – it is wrong – the possibilities of unintended bias and flawed inference emphasize the need to protect and insulate people and processes from computational misadventure.

6.5.4 Use of AI for predictive analytics in health care

Health care has always included and depended in part on predictions and prognoses and the use of predictive analytics. AI is one of the more recent tools for this purpose, and many possible benefits of prediction-based health care rely on use of AI. AI could also be used to assess an individual's risk of disease, which could be used for prevention of diseases, such as heart disease and diabetes. AI

could also assist health-care providers in predicting illness or major health events. For example, early studies with limited datasets indicated that AI could be used to diagnose Alzheimer disease years before symptoms appear [180].

Challenges to prediction in clinical care predate the emergence of AI and should not be attributed solely to AI techniques. Yet, various risks are associated with the use of AI to make predictions that affect patient care or influence the allocation of resources by a hospital or health-care system. Prediction technologies could be inaccurate because an AI technology bases its recommendations on an inference that optimizes markers of health rather than identifying an underlying patient need. An algorithm that predicts mortality from training data may have learnt that a patient who visits a chaplain is at increased risk of death [181].

While AI-based diagnosis is near term and its efficiency can be tested, thereby mitigating potential harm, efficacy and accuracy in long-term predictions may be more difficult or impossible to achieve. The risk of harm therefore increases dramatically, as predictions of limited reliability could affect an individual's health and well-being and result in unnecessary expenditure of scarce resources. For example, an AI-based mobile app developed by DeepMind to predict acute kidney failure produced two false-positive results for every correct result and therefore did not improve patient outcomes [182]. Even if the system identified some patients who required treatment, this benefit was cancelled out by overdiagnosis. Such false-positive results can harm patients if they persuade doctors to take riskier courses of action, such as prescription of a more potent, addictive drug, in response to the prediction.

Prediction-based health care, even if it is effective for diagnosis or accurate prediction of disease, may present significant risks of bias and discrimination for individuals because of a predisposition to certain health conditions [183], which could manifest itself in the workplace, health insurance or access to health-care resources.

The use of predictions throughout health care also raises ethical concern about informed consent and individual autonomy if predictions are shared with people who did not consent to surveillance, detection or use of predictive models to draw inferences about their future health status or to provide them with a "predictive diagnosis" that they did not request in advance. Such non-consensual misuse could include, for example, screening to predict psychotic episodes by analysis of speech patterns [184] or use of AI to identify individuals with tuberculosis who do not know their status (as described above) or at high risk of HIV infection and thus candidates for pre-exposure prophylaxis [185]. The Convention for the Protection of Human Rights and Dignity of the Human Being about the Application of Biology and Medicine (Oviedo Convention) [68] states that: "Everyone is entitled to know any information collected about his or her health. However, the wishes of individuals not to be so informed shall be observed."

Prediction-based technologies that are considered far more accurate or effective than older technologies could also challenge individual freedom of choice, even outside the doctor-patient relationship. Such use of AI, combined for example with "nudging", could transform an application for promoting healthy behaviour into a technology that could exert powerful control over the choices people make in their daily lives [105], because nudging and the many ways in which it can be done can be far more effective than sporadic interactions between a health-care provider and a patient. If AI predicts that an individual is at high risk of a certain disease, will that individual still have the right to engage in behaviour that increases the likelihood of the disease? Such restrictions on autonomy could be imposed by a doctor but also by an employer or insurer or directly by an AI application on a wearable device.

Thus, while the introduction of prediction-based algorithms is often well-intentioned, the challenges and problems associated with their use can cause more harm than benefit, as was a predictive algorithm for assessing the likelihood of pregnancy in adolescents in vulnerable populations (Box 5).

Box 5 – Challenges associated with a system for predicting adolescent pregnancy in Argentina

In 2017, the province of Salta, Argentina, signed an agreement with Microsoft to use AI to prevent adolescent pregnancy, a public health objective, and a tool to prevent school dropout. Microsoft used data for AI training collected by the local government from populations in vulnerable situations. The local authorities described the system [186] as:

> intelligent algorithms that identify characteristics in people that can lead to some of these problems [adolescent pregnancy and school dropout] and warn the government so that they can work on prevention.

The data processed by Microsoft servers were distributed globally. It was claimed that, on the basis of the data collected, the algorithm would predict whether an adolescent would become pregnant with 86% accuracy [187]. Once the partnership was publicized, however, it was challenged on technical grounds by local experts [188], for two reasons.

1. Testing of the algorithms for predicting adolescent pregnancy had significant methodological shortcomings. The training data used to build the predictive algorithm and the data used to evaluate the algorithm's accuracy were almost identical, which gave rise to an erroneous conclusion about the predictive accuracy of the system.
2. The type of data collected was inappropriate for ascertaining a future risk of pregnancy. The training data used were extracted from a survey of adolescents living in the province of Salta, which included personal information (e.g., age, ethnicity, country of origin), information about their environment (e.g., number of people in the household, whether they have hot water in the bathroom) and whether the person was pregnant at the time of the survey. These data were not appropriate for determining whether an individual would become pregnant in the future (e.g., within the ensuing 6 years), which would have required data collected 5 or 6 years before a pregnancy occurred. The collected data could be used at best only to determine whether an adolescent had been or was now pregnant.

The predictive algorithm was also inappropriate, as it provided predictions that were sensitive for adolescents without their (or their parents') consent, thereby undermining their privacy and autonomy. As the algorithm targeted individuals who were especially vulnerable, it was unlikely that they would have the opportunity to contest use of the interventions, and it could reinforce discriminatory attitudes and policies [189].

Despite the criticism and failings, the system continues to be used in at least two other countries (Brazil and Colombia) and in other provinces of Argentina [187]. The flaws in the algorithm would have been identified more easily if there had been greater transparency about the data sets used to train and evaluate the algorithm, the technical specifications and the hypothesis that guided the model's design [190].

This case study was written by Maria Paz Canales (Derechos Digitales).

6.5.5 Use of AI for prediction in drug discovery and clinical development

It is expected that machine-learning systems will be used to predict which drugs will be safe and effective and are best suited for human use. Machine learning may also be used to design drug combinations to optimize the use of promising AI or conventionally designed drug candidates. Such predictive models could allow pharmaceutical companies to take "regulatory shortcuts" and conduct fewer clinical trials and with fewer patient data. A possible benefit of AI may therefore be to accelerate the development of medicines and vaccines, especially for new diseases with pandemic potential for which there are ineffective or no medical countermeasures.

Such approaches can, however, carry risks if AI is used incorrectly or too aggressively. Predictive models are based on algorithms that must be assessed for accuracy, which may be difficult because of lack of transparency or explainability about how the algorithms function. Furthermore, reducing the number of trials or patients studied can raise concern that patients may be exposed to risks that were not identified by the algorithm.

6.6 Bias and discrimination associated with artificial intelligence

Societal bias and discrimination are often replicated by AI technologies, including those used in the criminal justice system, banking, human resources and the provision of public services. The different forms of discrimination and bias that a person or a group of people suffer because of identities such as gender, race and sexual orientation must be considered. Racial bias (in the USA and other countries) is affecting the performance of AI technologies for health (Box 6).

Box 6 – Discrimination and racial bias in AI technology

In a study published in Science in October 2019 [191], researchers found significant racial bias in an algorithm used widely in the US health-care system to guide health decisions. The algorithm is based on cost (rather than illness) as a proxy for needs; however, the US health-care system spent less money on Black than on white patients with the same level of need. Thus, the algorithm incorrectly assumed that white patients were sicker than equally sick Black patients. The researchers estimated that the racial bias reduced the number of Black patients receiving extra care by more than half.

This case highlights the importance of awareness of biases in AI and mitigating them from the onset to prevent discrimination (based on, e.g., race, gender, age or disability). Biases may be present not only in the algorithm but also, for example, in the data used to train the algorithm. Many other types of bias, such as contextual bias ([192], [193]), should be considered. Stakeholders, particularly AI programmers, should apply "ethics by design" and mitigate biases at the outset in developing a new AI technology for health [194].

Note – This case study was written by Marcelo Corrales Compagnucci (CeBIL Copenhagen), Sara Gerke (Harvard Law School) and Timo Minssen (CeBIL Copenhagen).

6.6.1 Bias in data

The data sets used to train AI models are biased, as many exclude girls and women, ethnic minorities, elderly people, rural communities and disadvantaged groups. In general, AI is biased towards the majority data set (the populations for which there are most data), so that, in unequal societies, AI may be biased towards the majority and place a minority population at a disadvantage. Such systematic biases, when enshrined in AI, can become normative biases and can exacerbate and fix (in the algorithm) existing disparities in health care [195]. Such bias is generally present in any inferential model based on pattern recognition. Thus, the human decisions that:

> comprise the data and shape the design of the algorithm [are] now hidden by the promise of neutrality and [have] the power to unjustly discriminate at a much larger scale than biased individuals [196].

Existing bias and established discrimination in health-care provision and the structures and practices of health care are captured in the data with which machine-learning models are trained and manifest in the recommendations made by AI-guided technologies. The consequence is that the recommendations will be irrelevant or inaccurate for the populations excluded from the data (Box 7), which is also the consequence of introducing an AI technology that is trained for use in one context into a different context.

Box 7 – AI technologies for detecting skin cancer exclude people of colour

Machine learning has outperformed dermatologists in detecting potentially cancerous skin lesions. As rates of skin cancer increase in many countries, AI technology would improve the ability of dermatologists to diagnose skin cancer. The data used to train one highly accurate machine-learning model are, however, for "fair-skinned" populations in Australia, Europe and the USA. Thus, while the technology assists in diagnosis, prevention and treatment of skin cancer in white and light-skinned individuals, the algorithm was neither appropriate nor relevant for people of colour, as it was not trained on images of these populations.

The inadequacy of the data on people of colour is due to several structural factors, including lack of medical professionals and of adequate information in communities of colour and economic barriers that prevent marginalized communities from seeking health care or participating in research that would allow such individuals to contribute data.

Another reason that such machine-learning models are not relevant for people of colour is that developers seek to bring new technologies to the market as quickly as possible. Even if their haste is guided by a desire to reduce avoidable morbidity and mortality, it can replicate existing racial and ethnic disparities, while a more deliberate, inclusive approach to design and development would identify and avoid biased outcomes.

Source: reference 197.

Such biases in data could also affect, for example, the use of AI for drug development. If an AI technology is based on a racially homogenous dataset, biomarkers that an AI technology identifies and that are responsive to a therapy may be appropriate only for the race or gender of the dataset and not for a more diverse population. In such cases, a drug that is approved may not be effective for the excluded population or may even be harmful to their health and well-being.

Data biases are also due to other factors. One is the digital divide. (See section 6.2.) Thus, women in LMIC are much less likely than men to have access to a mobile phone or mobile Internet; 327 million fewer women than men have access to mobile Internet [198]. Thus, women not only contribute fewer data to data sets used to train AI but are less likely to benefit from services. Another cause is unbalanced collection of data, even where the digital divide is not a factor. For example, genetic data tend to be collected disproportionately from people of European descent ([199], [200]). Furthermore, experimental and clinical studies tend to involve male experimental models or male subjects, resulting in neglect of sex-specific biological differences, although this gap may be closing slightly [201].

Biases can also emerge when certain individuals or communities choose not to provide data. Data on certain population subsets may be difficult to collect if collection requires expensive devices such as wearable monitors. As noted above, improving data collection from such communities or individuals, while it may improve the performance of AI, carries a risk of data colonialism. (See section 6.3.)

6.6.2 Biases related to who develops AI and the origin of the data on which AI is trained

Biases often depend on who funds and who designs an AI technology. AI-based technologies have tended to be developed by one demographic group and gender, increasing the likelihood of certain biases in the design. Thus, the first releases of the Apple Health Kit, which enabled specialized tracking of some health risks, did not include a menstrual cycle tracker, perhaps because there were no women on the development team [202].

Bias can also arise from insufficient diversity of the people who label data or validate an algorithm. To reduce bias, people with diverse ethnic and social backgrounds should be included, and a diverse team is necessary to recognize flaws in the design or functionality of the AI in validating algorithms to ensure lack of bias.

Bias may also be due to the origin of the data with which AI is designed and trained. It may not be possible to collect representative data if an AI technology is initially trained with data from local populations that have a different health profile from the populations in which the AI technology is used. Thus, an AI technology that is trained in one country and then used in a country with different characteristics may discriminate against, be ineffective or provide an incorrect diagnosis or prediction

for a population of a different race, ethnicity or body type. AI is often trained with local data to which a company or research organization has access but sold globally with no consideration of the inadequacy of the training data.

6.6.3 Bias in deployment

Bias can also be introduced during implementation of systems in real-world settings. If the diversity of the populations that may require use of an AI system, due to variations in age, disability, co-morbidities or poverty, has not been considered, an AI technology will discriminate against or work improperly for these populations. Such bias may manifest itself at the workplace, in health insurance or in access to health-care resources, benefits and other opportunities. As AI is designed predominantly in HIC, there may be significant misunderstanding of how it should be deployed in LMIC, including the discriminatory impact (or worse) or that it cannot be used for certain populations.

6.7 Risks of artificial intelligence technologies to safety and cybersecurity

This section discusses several risks for safety and cybersecurity associated with use of AI technologies for health, which may be generalized to the use of many computing technologies for health care – past and present.

6.7.1 Safety of AI technologies

Patient safety could be at risk from use of AI that may not be foreseen during regulatory review of the technology for approval. Errors in AI systems, including incorrect recommendations (e.g., which drug to use, which of two sick patients to treat) and recommendations based on false-negative or false-positive results, can cause injury to a patient [159] or a group of people with the same health condition. Model resilience, or how an AI technology performs over time, is a related risk. Health-care providers also commit errors of judgement and other human errors, but the risk with AI is that such an error, if fixed in an algorithm, could cause irreparable harm to thousands of people in a short time if the technology is used widely [159]. Furthermore, the psychological burden and stress of such errors is borne by the providers who operate such technologies.

An AI application, like any information technology system, could also provide the wrong guidance if it has code errors due to human programming mistakes. For example, the United Kingdom NHS COVID-19 application, which was designed to notify individuals to self-isolate if exposed, was programmed incorrectly [203]. Thus, a user of the application had to be next to a highly infectious patient five times longer than that considered risky by the NHS before being instructed to self-isolate. Although up to 19 million people downloaded the application, a "shockingly low" number of people were told to isolate, thereby exposing themselves and others to risks of COVID-19 infection [203].

It is also possible that a developer (or an entity that funds or directs the design of AI technology) designs an AI technology unethically, to optimize an outcome that would generate profits for the provider or conceal certain practices. The design might in fact be more accurate than another modelling technique but generate unmerited sales revenue. Malicious design has affected other sectors, such as the automobile sector, in which algorithms used to measure emissions were programmed to conceal the true emissions profile of a major car manufacturer [204].

Use of computers carries an inherent risk of flaws in safety due to insufficient attention to minimizing risk in the design of machines and also to flaws in the computer code and associated bugs and glitches. Injuries and deaths due to such flaws and breakdowns are underreported, and there are no official figures and few large-scale studies. In one study in the United Kingdom, for instance, it was estimated that up to 2000 deaths a year may be due to computer errors and flaws and that it is an "unnoticed killer" [205].

6.7.2 Cybersecurity

As health-care systems become increasingly dependent on AI, these technologies may be expected to be targeted for malicious attacks and hacking in order to shut down certain systems, to manipulate

the data used for training the algorithm, thereby changing its performance and recommendations, or to "kidnap" data for ransom [181]. AI developers might be targeted in "spear-fishing" attacks and by hacking, which could allow an attacker to modify an algorithm without the knowledge of the developer.

An algorithm, especially one that runs independently of human oversight, could be hacked to generate revenue for certain recipients, and large sums are at stake: total spending on health care globally was US$ 7.8 trillion in 2017, or about 10% of global gross domestic product [206]. The United Kingdom Information Commission Office noted that cyberattacks on the health sector are the most frequent [207]. Breaches of health data, which are some of the most sensitive data about individuals, could harm privacy and dignity and the broader exercise of human rights. A study in 2013 showed that four anonymized data points are sufficient for unique identification of an individual with 95% accuracy [208]. Measures to avoid such breaches, which can be broadly categorized as infrastructural or algorithmic, are improving, although no defence is 100% effective and new defences can be broken as quickly as they are proposed [181].

6.8 Impacts of artificial intelligence on labour and employment in health and medicine

The impact of AI on the health workforce is viewed with equal optimism and pessimism. It is perhaps less contested that nearly all jobs in health care will require a minimum level of digital and technological proficiency. The Topol Review: Preparing the health workforce to deliver the digital future [24], concluded that, within two decades, 90% of all jobs in the United Kingdom's NHS will require digital skills, including navigating the "data-rich" health-care environment, and also digital and genomics literacy. The requirement for digital literacy will not be limited to clinical care (although this section concentrates on clinical staff) but extends to health-care workers in public health, surveillance, the environment, prevention, protection, education, awareness, diet, nutrition and all the other social determinants of health that can be supported by AI. All health workers in these areas will have to be trained and retrained in use of AI to support and facilitate their tasks.

Optimistic views include that in which AI will automate and thus reduce the burden of routine tasks on clinicians and allow them to focus on more challenging work and to engage with patients. It could also empower doctors to work in more areas and provide support in areas in which technology can be used for clinical decision-making. It is expected that digitization of health care and the introduction of AI technologies will create numerous new jobs in health care, such as software development, health-care systems analysis and training in the use of AI for health care and medicine. The last may include three types of jobs: trainers, or people that can evaluate and stress-test AI technologies; explainers, or those who can explain how and why an algorithm can be trusted; and "sustainers", or those who monitor behaviour and identify unintended consequences of AI systems [181].

AI could also extend one of the scarcest resources in health-care systems – the time that doctors and nurses have to attend to patients. If doctors and nurses can hand over repetitive or administrative tasks to AI-supported technologies and therefore spend less time on "routine care cases", they would have more time to attend to more urgent, complex or rare cases and to improve the overall quality of care offered to patients [24]. In some cases, however, as AI is being integrated into health-care systems as secondary medical support, during what could best be described as a transition period, AI may increase the tasks and add to the workload of doctors and nurses.

Telemedicine has been used to extend health-care provision to people in remote areas and to refugees and other underserved populations that otherwise lack appropriate medical advice [205]. Yet, AI and its use in telemedicine could create inequitable access to health-care services (in particular to health-care personnel), for instance when people in rural areas or low-income countries have to make do with greater access to AI-based services and telemedicine [181] while individuals in HIC and urban areas continue to benefit from in-person care.

Furthermore, health-care workers who already have to absorb large amounts of information to meet standards of care may regularly require new competence in the use of AI-supported technologies in

everyday practice, and competence may have to evolve rapidly as the uptake of AI accelerates. Such continuing education may be neither available nor accessible to all health-care workers, although efforts are under way to improve digital literacy and training that includes use of AI and other health information technologies. (See section 7.2.)

Even as health-care workers have to obtain new competence, the use of AI to augment and possibly replace the daily tasks of health-care workers and physicians could also remove the need for maintaining certain skills, such as the ability to read an X-ray. At some point, physicians may be unable to conduct such a task without the assistance of a computer, and AI systems will have to be "trained" to use the repository of medical knowledge that was the domain of human providers [159]. Such dependence on AI systems could erode independent human judgement and, in the worst-case scenario, could leave providers and patients incapable of acting if an AI system fails or is compromised [159]. There should therefore be robust plans to provide back-up if technology systems fail or are breached.

Another concern is that AI will automate many of the jobs and tasks of health-care personnel, resulting in significant loss of jobs in nearly every part of the health workforce, including certain types of doctors. AI has already replaced many jobs in other industries, reduced the total number of people required for certain roles or created the expectation that many jobs will be lost (e.g., up to 35% of all jobs in the United Kingdom) [210].

In many countries, however, health care is not an industry but a core government function, so that administrators will not replace health-care workers with technology. Many countries, with high, middle or low income, are in fact facing shortages of health-care workers. WHO has estimated that, by 2030, there will be a shortage of 18 million health workers, mostly in low- and low- to middle-income countries [211]. AI may provide a means to bridge the gap between the workforce ideally available to provide appropriate health care and what exists.

Other scenarios have been envisaged with the arrival of AI. One predicts that a decision to use AI will cause short-term instability, with many job losses in certain areas even as overall employment increases with the creation of new jobs, resulting in unemployment for those who may not be able to retrain for the new roles. In another scenario, job losses will not materialize, either because clinicians or health-care workers will fulfil other roles or because these technologies will be fully integrated only over a long time, during which other roles for health-care workers and clinicians will emerge, such as labelling data or designing and evaluating AI technologies [210].

Even if AI does not displace clinicians, it could make doctors' jobs less secure and stable. One trend has been the "Uberization" of health care, in which AI facilitates the creation of health-care platforms on which contractors, including drivers, temporary workers, nurses, physician assistants and even doctors, work on demand ([103], [211]). During the past decade, health care and education have seen the fastest growth of "gig workers", who work on a temporary basis with no stability of employment [103]. While this provides more flexible services, it could also sever relationships between patients and health-care givers and create insecurity for certain types of health workers. Such a trend may not occur in countries with either greater labour protection for its health workforce, such that labour shortages provide health-care workers with negotiating power, or in which AI is not used to reorganize health care but to reduce the workload.

With increasing use of AI, the nature of medical practice and health-care provision will fundamentally change. As noted above, it could provide health-care workers with more time to care for patients or it could, if patients interact more frequently and directly with AI, result in doctors spending less time in direct contact with patients and more time in administering technology, analysing data and learning how to use new technologies. If introduction of AI is not effectively managed, physicians could become dissatisfied and even leave medical practice [213].

6.9 Challenges in commercialization of artificial intelligence for health care

There are various ethical challenges to the practices of the largest technology firms in the field of AI for health, although some of the concerns also apply to mid-size firms and start-ups. The use of AI for health has been pushed by companies – from small start-up firms to large technology companies – mainly by significant advocacy and investment. Those who support a growing role for these companies expect that they will be able to marshal their capital, in-house expertise, computing resources and data to identify and build novel applications to support providers and health systems. During the COVID-19 pandemic, many companies have sought to provide services and products for the response, many of which are linked to forms of public health surveillance [214]. This has raised a number of ethical and legal concerns, which are discussed throughout this document.

Some services already widely used in health are for "back-office" functions and for managing health-care systems. Some of the companies involved in development of technology, such as the pharmaceutical and medical device industries, are integrating AI into their processes and products, and insurance firms are using AI for assessing risk or even automating the provision of insurance, which might raise ethical concerns with respect to algorithmic decision-making.

A prominent use of AI for health care is to support diagnosis, treatment, monitoring and adherence to treatment. Such applications could have benefits for health-care systems; however, many concerns have emerged during the past as more technology firms, and especially the largest firms, have entered the health-care field.

A general problem is lack of transparency. While many firms know much about their users, their users, civil society and regulators know little about the activities of the firms, including how they (and governments) operate in PPPs, which have a significant impact on the public interest [215]. (See section 9.3.) Their practices remain hidden partly because of commercial secrecy agreements or the lack of general obligations for transparent practices, including the role these firms play in health care and the data that are collected and used to train and validate an AI algorithm. Without transparency (and accountability), these firms have little incentive to act in a way that does not cross certain ethical boundaries or to disclose deeper problems in their technology, data or models [215]. Many companies prefer to keep their algorithmic models proprietary and secret, as full transparency could lead to criticism of both the technology and the company [216].

A second broad concern is that the overall business model of the largest technology firms includes both aggressive collection and use of data to make their technologies effective and use of surplus data for commercial practices, considered by Professor Shoshana Zuboff as "surveillance capitalism" [125]. Thus, during the past decade, there have been several examples of large technology firms using large datasets of sensitive health information in developing AI technologies for health care ([129], [217]). While such health data may have been acquired and used to develop useful AI technologies for health, the data were not acquired with the explicit consent of those who provided them, the benefits of the data for these firms may be far in excess of what was required to deliver the product, and the firms may not provide equal benefits to the population that generated the data in the first place.

Such acquisition of sensitive health information can give rise to legal concern. First, even if the data are anonymized by the firm that acquires them, the company would be able to combine data and de-anonymize relevant data sets from the amount of information it already has from other sources [147]. Secondly, several large technology firms have been accused and even fined for mishandling data [218], and this concern may be heightened for firms that acquire often-sensitive health data. Thirdly, as firms continue to accumulate large amounts of data, this can introduce anti-trust concerns (although it may not lead to regulatory enforcement [219]), related to the growing market power of such companies, including barriers to smaller companies that may wish to enter an AI market [220].

An additional concern is the growing power that some companies may exert over the development, deployment and use of AI for health (including drug development) and the extent to which

corporations exert power and influence over individuals and governments and over both AI technology and the health-care market. Data, computing power, human resources and technology can be concentrated within a few companies, and technology can be owned either legally (IP protection) or because the size of a company's platform results in a monopoly. Monopoly power can concentrate decision-making in the hands of a few individuals and companies, which can act as gatekeepers of certain products and services [221] and reduce competition, which could eventually translate into higher prices for goods and services, less consumer protection or less innovation.

While the growing role of large companies in the USA, such as Google, Facebook and Amazon, in the development and provision of AI for health care has been under scrutiny, large technology companies in China and other Asian countries are playing a similar role in health through such services and technologies. They include Ping An, Tencent, Baidu and Alibaba, which are both building their own technology platforms and collaborating with user platforms such as WeChat to reach millions of people in China [176]. Tencent, for example, is investing in at least three main areas of health: AI-based technologies to assist in diagnosis and treatment, a "smart hospital" to provide a web of online services and data connectivity through a smart health card (which itself raises concern about data privacy and use; see above) and a "medipedia" to provide health information to users online [222]. Alibaba is working with hospitals to predict patient demand in order to allocate health-care personnel and developing AI-assisted diagnostic tools for radiology [176].

Such power and control of the market by large firms may be part of a 'first-mover' advantage that several large firms may eventually earn through their entry into AI for health. Even if the data used by a firm (for example, data from a public health system) could be used by others, other firms might be discouraged or unable to replicate use of such data for a similar purpose, especially if another company has already done so [215]. Such power also means that the rules set by certain companies can force even the largest and wealthiest governments to change course. For example, during the COVID-19 pandemic, Google and Apple introduced a technical standard for where and how data should be stored in proximity-tracking applications that differed from the approach preferred by the governments of several HIC, which resulted in at least one government changing the technical design of its proximity-tracking application to comply with the technical standards of these two companies. Although the approach of these companies may have been consistent with privacy considerations, the wider concern is that these firms, by controlling the infrastructure with which such applications operate, can force governments to adopt a technical standard that is inconsistent with its own public policy and public health objectives [223].

When most data, health analytics and algorithms are managed by large technology companies, it will be increasingly likely that those companies will govern decisions that should be taken by individuals, societies and governments, because of their control and power over the resources and information that underpins the digital economy [124]. This power imbalance also affects people who should be treated equitably by their governments or at least, if treated unfairly, can hold their governments accountable if inequity arises. Without a strong government role, companies might ignore the needs of individuals, particularly those at the margins of society and the global economy [179].

Stringent oversight by governments and good governance are essential in this sector. (See section 9.3 on private sector governance.) Oversight mechanisms could be integrated into PPPs. If such partnerships are not carefully designed, they can lead to misappropriation of resources (usually patient data) or conflicts of interest in decision-making in such partnerships or could forestall or limit the use of regulation to protect the public interest when necessary ([215], [216]).

6.10 Artificial intelligence and climate change

Use of deep learning models in AI has been scrutinized for its impact on climate change. Researchers at the University of Massachusetts Amherst, USA, found that the emissions associated with training a single "big language" model were equal to approximately 300 000 kg of carbon dioxide or 125 round-trip flights between New York City and Beijing [224]. A single training session for another

deep-learning model, GTP-3, requires energy equivalent to the annual consumption of 126 Danish homes and creates a carbon footprint equivalent to travelling 700 000 km by car [225]. All the infrastructure required to support use of AI has an additional carbon cost [225].

WHO considers climate change to be an urgent, global health challenge that requires prioritized action now and in the decades to come. Between 2030 and 2050, climate change is expected to cause approximately 250 000 additional deaths per year from malnutrition, malaria, diarrhoea and heat stress alone. The cost of direct damage to health by 2030 is estimated to be US$ 2-4 billion per year. Areas with weak health infrastructure – most in developing countries – will be the least able to cope without assistance to prepare and respond [226].

Reducing emissions of greenhouse gases through better transport, food and choices of energy, particularly reducing air pollution, results in better health [226]. Extending the use of AI for health and in other sectors of the global economy could, however, contribute directly to dangerous climate change and poor health outcomes, especially of marginalized populations. Thus, the growing success and benefits for health outcomes of AI, which will predominate in HIC, would be directly linked to increased carbon emissions and negative consequences in low-income countries. AI technologies, for health and other uses, should therefore be designed and evaluated to minimize carbon emissions, such as by using smaller, more carefully curated data sets, which could also potentially improve the accuracy of AI models [227]. Otherwise, the growing use of AI might have to be balanced against its impact on carbon emissions.

7 Building an ethical approach to use of artificial intelligence for health

This section addresses how measures other than law and policy can ensure that AI improves human health and well-being.

7.1 Ethical, transparent design of technologies

Although technology designers and developers play critical roles in designing AI tools for use in health, there are no procedures for credentialing or licensing such as those required for health-care workers. In the absence of formal qualifications for ethics in the AI field, it is not enough merely to call for personal adherence to values such as reproducibility, transparency, fairness and human dignity.

New approaches to software engineering in the past decade move beyond an appeal to abstract moral values, and improvements in design methods are not merely upgraded programming techniques. Methods for designing AI technologies that include moral values in health and other sectors have been proposed to support effective, systematic, transparent integration of ethical values. Such values in design have also been codified legally; for example, the GDPR includes specific obligations to include privacy by design and by default.

One approach to integrating ethics and human rights standards is "Design for values", a paradigm for basing design on the values of human dignity, freedom, equality and solidarity (Box 8) and for construing them as non-functional requirements [228]. This requires not a solutions-oriented approach but instead a process-oriented approach that satisfies stakeholder needs in conformity with the moral and social values embodied by human rights.

Box 8 – Design for values [229]

"Design for values" is explicit transposition of moral and social values into context-dependent design requirements. It is an umbrella term for several pioneering methods, such as value-sensitive design, values in design and participatory design. Design for values presents a roadmap for stakeholders to translate human rights into context-dependent design requirements through a structured, inclusive, transparent process, such that abstract values are translated into design requirements and norms (properties that a technology should have to ensure certain values), and the norms then become a socio-technical design requirement. The process of identifying design requirements permits all stakeholders, including individuals affected by the technology, users, engineers, field experts and legal practitioners, to debate design choices and identify the advantages and shortcomings of each choice.

Thus, a value such as privacy can be interpreted through certain norms, such as informed consent, right to erasure and confidentiality. These norms can then be converted by discussion and consultation into design requirements, such as positive opt-in (a means of ensuring informed consent) or homomorphic encryption techniques to assure confidentiality. Other techniques for safeguarding privacy, such as *k*-anonymity, differential privacy and coarse graining through clustering, could also be selected through consultation.

Ethical design can also be applied to the socio-technical systems in which algorithms are developed, which comprise the ensemble of software, data, methods, procedure, personnel, protocols, laws, norms, incentive structures and institutional frameworks. All are brought together to ensure that products and services provide ethical outcomes for society and its health-care systems.

More generally, ethical and transparent design of AI technologies should be ensured by prioritizing inclusivity in processes and methods ([230], [231]). Consideration of inclusivity when designing and developing an AI technology can overcome barriers to equitable use of the technology in health associated with geography, gender, age, culture, religion or language.

Three approaches for promoting inclusivity are the following.

– *Citizen science:* Citizen science is defined by the Alan Turing Institute as the direct contribution of non-professional scientists to scientific research, for instance, by contributing data or performing tasks [232]. Citizen science not only helps the public to understand a particular study or technology that may affect them personally but also ensures that the public is involved in research, discussions and tool-building. This ensures respectful co-creation of AI technologies that reduces the distance between the researcher or programmer and the individuals who the technology is intended to serve.

– *Open-source software*: Transparency and participation can be increased by the use of open-source software for the underlying design of an AI technology or making the source code of the software publicly available. Open-source software is open to both contributions and feedback, which allows users to understand how the system works, to identify potential issues and to extend and adapt the software. Open-source software design must be accessible and welcoming, and the content should allow greater engagement and transparency.

– *Increased diversity:* Too often, efforts to increase the diversity of AI technologies involve increasing the diversity of the data on which they are based. Although this is necessary, it is not sufficient and might even amplify any biases inherent in the design. Minimizing and identifying potential biases requires greater involvement of people who are familiar with the nature of potential biases, contexts and regulations throughout software development, from its design to consultation with stakeholders, labelling of data, testing and deployment.

Toolkits can be useful for providing concrete guidance to technology designers who wish to integrate ethical considerations into their work. Software developer kits can provide guidelines that include a code of ethics, with specific guidelines for health. Such kits could indicate, for example, how to manage data, including collection, de-identification and aggregation, and how to safeguard the destination of data.

Kits have also been developed to facilitate certain ethical (and increasingly legal) requirements, such as the Sage Bionetworks toolkit for the elements of informed consent [233]. The toolkit provides use cases to explain its approach to informed consent, including eConsent, examples of how it should be put into practice, a checklist to ensure that programmers have considered all the necessary questions and additional resources.

With the proliferation of use of AI for health, the emergence of more not-for-profit AI developers would be beneficial. Such developers, who are not constrained by internal or external revenue targets, can adhere to ethical principles and values more readily than private developers. Not-for-profit developers may include treatment providers, hospital systems and charities. They could emulate the many partnerships for not-for-profit product development that have been formed during the past two decades in the development of new medicines, diagnostics and vaccines. The partnerships are often with the public and private sectors and focus on neglected populations while ensuring affordability and access to all. A not-for-profit developer could address all areas of health but particularly areas of neglect, while ensuring that their technologies adhere to ethical values such as privacy, transparency and avoidance of bias.

Putting prediction to good use

Use of AI for prognosis will allow assessment of the relative risk of disease and predict illness. There are, however, several risks and challenges with the use of predictive analytics, including concern about the accuracy of the predictions and that prediction of a negative outcome could affect an individual's autonomy and well-being.

In public health, predictive analytics can forecast major health events, including outbreaks, before they occur. For example, before the COVID-19 pandemic, WHO was developing EPI-BRAIN, a global platform that will allow experts in data and public health to analyse large datasets for use in emergency preparedness and response [234]. It would allow forecasting and early detection of threats of infection and their impact on the basis of scenarios, simulation exercises and insights to improve coordinated decision-making and response.

Ethical, transparent design allows governments and international health agencies, such as WHO, to encourage the development of AI technologies for predictive analytics to assist and augment decision-making by providers and policy-makers. Such technologies must adhere to ethical standards and human rights obligations, should be open to improvement and should be available for adaptation and use by governments and providers on a non-exclusive basis.

Recommendations

1. Potential end-users and all direct and indirect stakeholders should be engaged from the early stages of AI development in structured, inclusive, transparent design and given opportunities to raise ethical issues, voice concerns and provide input for the AI application under consideration. Relevant ethical considerations should inform the design and translation of moral values into specific context-dependent design requirements.
2. Designers and other stakeholders should ensure that AI systems are designed to perform well-defined tasks with the accuracy and reliability necessary to improve the capacity of health systems and advance patient interests. Designers and other stakeholders should also be able to predict and understand potential secondary outcomes.
3. Designers should ensure that stakeholders have sufficient understanding of the task that an AI system is designed to perform, the conditions necessary to ensure that it can perform that task safely and effectively and conditions that might degrade system performance.
4. The procedures that designers use to "design for values" should be informed and updated by the consensus principles stated in this document, best practices (e.g., privacy preserving technologies and techniques), standards of ethics by design, evolving professional norms (transparency of access to codes, processes that allow verification and inclusion).

5. Continuing education and training programmes should be available to designers and developers to ensure that they integrate evolving ethical considerations into design processes and choices. The establishment of formal accreditation procedures could ensure that designers and developers abide by ethical principles similar to those required of health-care workers.

7.2 Engagement and role of the public and demonstration of trustworthiness to providers and patients

Effective use of AI for health will require building the trust of the public, providers and patients. Social license requires hard-fought efforts that can be surrendered quickly if AI technologies are introduced without due care for the perspectives of those affected by its use. Public engagement and dialogue are means to ensure that use of AI for health care meets certain core societal expectations and greater trust and acceptance. Public dialogue also allows ascertainment of society's views, as far as possible, on the ethical dimensions of AI, its design and uses.

A critical issue of public concern, discussed throughout this publication, is the collection and use of patient data for AI and other applications. In the United Kingdom, these concerns have been addressed in public debate and dialogue. Health Data Research, which collects health data and makes it available to public and private entities for health-related applications of AI,[7] has used public engagement, including with the Wellcome Trust's initiative, Understanding Patient Data [236]. Workshops held as part of the initiative provided a forum for participants to discuss their expectations and concerns about use of patient data in AI and other applications. Before these workshops, 18% of participants considered it acceptable to share anonymized patient data with commercial organizations for reasons other than direct care; after the workshops, the proportion had increased to 45% [237]. Individuals who expressed positive views considered that contributing data was a value exchange, with a societal benefit, and wanted the NHS to benefit from their data. They also considered it acceptable for commercial companies to have access to their data, provided that the benefit returned to the public and that the NHS administered the data for the public benefit.

The United Kingdom Academy of Medical Sciences found at its meetings and workshops [238] that:

> ongoing engagement with patients, the public and healthcare professionals, including via co-creation, will be critical to ensuring new AI technologies respond to clinical unmet need, are fit for purpose, and are successfully deployed, adopted and used.

The Academy conducted a public dialogue on the "data-driven future" to understand awareness, expectations, aspirations and concerns about future technologies that would require patient data to be accessed, analysed or linked for clinical diagnosis and management [239]. The respondents considered that any new use of data must have a proven social benefit and that an appropriate organization (such as the government or the NHS) should oversee the data and administer it for the public benefit.

Steps must be taken to build the trust of providers and patients who will increasingly rely on AI for routine clinical decision-making. The willingness of patients to rely on AI may sometimes be much lower than expected. For example, in a study conducted by HSBC Bank [240], only 8% of the respondents surveyed said that they would trust a machine offering mortgage advice, while 41% said they would trust a mortgage broker. Lack of wider trust could create significant divisions in a health-care system, in which, for example, older patients might be unwilling to adapt to and use new AI technologies, while younger patients might be more amenable [155].

With such a low level of trust, scandals that emerge from use of AI for health care and undermine patients' economic, personal or physical security could be fatal. After the Cambridge Analytica

---

[7] Presentation by Dr Andrew Morris, Health Data Research United Kingdom, 3 October 2019 to the WHO working group on ethics and governance of AI for health.

scandal in 2019, an estimated 15% of Facebook users surveyed indicated they would reduce their use of the social networking site. Trust could be eroded even more quickly and severely in the domain of health care if similar scandals or abuses of trust emerged into public discourse, destroying public trust overnight [158].

One means of mitigating and managing risk would be to allow health-care providers and developers to test a new AI product or service in a "live environment" in a testing facility, with safeguards and oversight to protect the health system from any risks or unintended consequences. Testing facilities could allow assessment, certification and validation of AI. In limited circumstances, testing facilities could build a "regulatory sandbox" [241], which might, however, be appropriate only in countries in which new health-care products and services and their specifications are subject to formal regulation and to data protection regulations [242]. Examples of the use of regulatory sandboxes are the United Kingdom's Care Quality Commission and by the Singapore Government to test new digital health models [242].

A second approach to building trust and facilitating a "graceful transition" of health care is to redesign training programmes for the health workforce (Box 9) and to improve general education [243]. Improvements in general education would include primary education in science, technology and mathematics.

Box 9 – Supporting health workers in the use AI technologies, including through education and training

Medical professionals and health-care workers should receive sufficient technical, managerial and administrative support, capacity-building, regulatory protection (when appropriate) and training in the many uses of AI technologies and their advantages and in navigating the ethical challenges of AI [244]. With regard to education and training, AI curricula should be seamlessly integrated into existing programmes [244]. Curricula should be updated regularly, as AI is evolving continuously. Some members of the health-care profession will require training in basic use of computers before they adapt to use of AI. All health-care professionals will require a certain level of digital literacy, defined in the Topol review as "those digital capabilities that fit someone for living, learning, working, participating and thriving in a digital society" [24].

Physicians and nurses will also require a wider range of competence to apply AI in clinical practice, including better understanding of mathematical concepts, the fundamentals of AI, data science, health data provenance, curation, integration and governance [24], and also of the ethical and legal issues associated with the use of AI for health. Such measures (including training) will be necessary to combine and analyse data from many sources adequately, supervise AI tools and detect inaccurate performance of AI [244]. Good support and training will ensure that health-care workers and physicians, for example, can avoid common pitfalls such as automation bias when using AI technologies. Eventually, the knowledge, skills and capabilities required of health workers may be defined by professional and statutory regulatory bodies in collaboration with practitioners and educators [24].

Significant changes may be made to medical education. Rather than rote memorization, which has been the hallmark of medical training, medical students might instead build and refine their competence for communication and negotiation, emotional intelligence, the ability to resolve ethical dilemmas and proficient use of computers. Medical training programmes will therefore require new educators who can teach these concepts and skills [24].

A third approach, the use of human warranty, is discussed earlier in this document (section 5), whereby developers of AI technologies work directly with providers and patients in patient and clinical evaluation at critical points in the development and deployment of the technologies. Human warranty can ensure meaningful public consultation and debate [101].

Recommendations

1. The public should be engaged in the development of AI for health in order to understand forms of data sharing and use, to comment on the forms of AI that are socially and culturally acceptable and to fully express their concerns and expectations. Further, the general public's literacy in AI technology should be improved to enable them to determine which AI technologies are acceptable.
2. Training and continuing education programmes should be available to assist health-care professionals in understanding and adapting to use of AI, learning about its benefits and risks and understanding the ethical issues raised in their use.

7.3 Impact assessment

An impact assessment is used to predict the consequences of a current or proposed action, policy, law, regulation or, as in the case of use of AI for health, a new technology or service. Impact assessments can provide both technical information on possible consequences and risks (both positive and negative) and improve decision-making, transparency and participation of the public in decision-making and introduce a framework for appropriate follow-up and measurement. Such assessment might be especially important for the use of AI, as an AI technology can change over time [245]. Impact assessments can also be used to determine whether a technology will respect or undermine ethical principles and human rights obligations, including privacy and non-discrimination. Several types of impact assessment for the use of AI for health have been proposed or used, which could be considered by governments, companies and providers.

Businesses that design and introduce AI technologies for health have a particular obligation to conduct impact assessments, including on human rights. The UN Guiding Principles on Business and Human Rights of the United Nations Office of the High Commissioner for Human Rights establish corporate responsibility to respect human rights, including for companies to conduct due diligence to identify, avoid, mitigate and remedy impact on human rights for which they are responsible or indirectly involved [246]. Although the UN Guiding Principles do not require businesses to conduct human rights impact assessments, such an assessment can help companies to meet their obligations.

Impact assessments allow identification, understanding, assessment and mitigation of the adverse effects of business projects or activities on human rights [247]. Although such assessments are relatively new, their use has increased, including for the deployment of AI. The United Nations Special Rapporteur on Freedom of Expression noted [3].

> Human rights impact assessments and public consultations should be carried out during the design and deployment of new AI systems, including the deployment of existing AI systems in new global markets.

Human rights impact assessments have also been recognized in national laws as an obligation of companies. For example, the French Government enacted a law on "duty of vigilance" that requires parent companies to identify and prevent adverse impacts on human rights and the environment resulting from their activities, from the activities of companies that they control and from the activities of the subcontractors and suppliers with which they have commercial relations [248]. Furthermore, a EU Directive may require all companies with headquarters in Europe to conduct human rights due diligence, although the discussions will be completed only in 2021 [249].

Other types of impact assessment have been either proposed or implemented. One approach is an "ethical impact assessment" to identify the impacts of AI on human rights, including in vulnerable groups, labour rights, environmental rights and their ethical and social implications. A second approach, proposed by the AI Now Institute, is an "algorithmic impact assessment" for public agencies, as a "practical framework to assess automated decision systems and to ensure public accountability" [250]. Such assessments would be both for affected communities to obtain information on how automated decision systems function and to determine whether they are

acceptable and also for governments to assess how the systems are used, whether they have disparate impacts in particular on the basis of gender, race or another dimension and how to hold the systems accountable. This could be useful for governments as they turn to algorithmic decision-making for large- and small-scale health-care decisions.

Several laws have been proposed or implemented that require impact assessments, including for the use of AI for health. In 2019, two senators in the USA co-sponsored the "Algorithmic Accountability Act", which would require companies to study and adjust flawed algorithms that result in inaccurate, unfair, biased or discriminatory decisions that would affect people in the USA [251]. It would also require companies, with enforcement by the US Federal Trade Commission, to "reasonably address" the results of such assessments, including algorithmic decisions that affect health. Such assessments would be made only for "high-risk" decisions, which would include health information or genetic data or decisions or analyses of sensitive aspects of individual lives, including their health and behaviour. The act has, however, only been proposed and is not enacted [251].

A separate proposal under the proposed legislation would require companies to conduct "data protection impact assessments" for high-risk information systems, such as those that store or use personal information, including health information. This would mirror the impact assessment required by law under the EU GDPR, which requires companies to conduct 'data impact assessments" of the risks of data processing operations to the "rights and freedoms of natural persons" and their impact on the protection of personal data [252].

Recommendations

1. Governments should enact laws and policies that require government agencies and companies to conduct impact assessments of AI technologies, which should address ethics, human rights, safety and data protection, throughout the life-cycle of an AI system.
2. Companies and developers should conduct impact assessments as per the UN Guiding Principles on Business and Human Rights, even if governments have not mandated them.
3. Impact assessments should be audited by an independent third party before and after introduction of an AI technology and published.

7.4 Research agenda for ethical use of artificial intelligence for health care

In a fast-moving field such as the use of AI for health, there are many unresolved technical and operational questions on how best to use AI. Use of AI also generates ethical quandaries. Each new application or use of AI raises opportunities and challenges that should be addressed before widespread adoption. This has been the case for the proliferation and deployment of new AI technologies during the COVID-19 pandemic.

Suggested areas of research to address emerging issues and challenges

Some ethical concerns require research to substantiate and explain the challenges. Approaches to addressing concerns should be tested and validated with research, such as on computer science or on the consequences of using AI for a particular medical need or target population. Research on each of these topics should include consideration of different countries, cultures and types of health-care systems. Pertinent research questions include the following.

– For what needs and gaps identified by health-care workers and patients could AI play a role in ensuring the delivery of equitable care?
– How is AI changing the relationships between health-care workers and patients? Do these technologies allow providers to spend more "quality" time with patients, or do they make care less humane? Do specific contextual factors improve or undermine the quality of care?

– What are that attitudes of health-care workers and patients towards the use of AI? Do they find these technologies acceptable? Do their attitudes depend on the type of intervention, the location of the intervention or current acceptance of these technologies both in the health-care system and in society?
– Has the introduction and use of AI for health exacerbated the digital divide? Or does AI, with telemedicine, reduce the gap in access to care and ensure equitable access to high-quality care, irrespective of geography and other demographic factors?
– How best can providers and programmers address any biases that will manifest in applications? What are the barriers to addressing biases?
– What method should be used to assess whether AI is more cost-effective and appropriate than existing or "low-technology" solutions in LMIC? How should governments and providers assess fair resource allocation for existing interventions and new technologies?
– Can ethical design be applied specifically to AI technologies for health?

8 Liability regimes for artificial intelligence for health

Although the performance of machine-learning algorithms is improving, there will still be errors and mistakes, for example because an algorithm has been trained with incomplete or inappropriate data, programming mistakes or security flaws. Even AI technologies designed with well-curated data and an appropriate algorithm could harm an individual. While AI technologies may be safe in practice, unforeseeable risks are likely [253].

Lawmakers and regulators should ensure that rules and frameworks for safety are applicable to the use of AI technologies for health care and that they are proactively integrated into the design and deployment of AI-guided technologies. Updated liability rules for the use of AI in clinical care and medicine should at least include the same standards and damages already applied to health care. It is possible that reliance on AI technologies and the risks they may pose require additional obligations and damages. This section addresses how liability regimes could evolve, approaches to compensation, specific considerations for LMIC and the role of international institutions and organizations. It does not address liability that may arise from data processing.

8.1 Liability for use of artificial intelligence in clinical care

Use of AI to support or augment clinical decision-making raises several questions. Should doctors be held at fault if they follow the suggestion of an AI technology that results in a medical error or if they ignore a suggestion that would have avoided morbidity or mortality? The answers to these questions depend largely on other choices, such as the types of behaviour encouraged or discouraged by a legal system and the standard of care as use of AI in clinical practice becomes more common.

Another choice is whether liability rules should encourage clinicians to rely upon AI to inform and confirm their clinical judgement or to deviate from their own judgement if an algorithm arrives at an unexpected conclusion. If liability rules penalize health-care providers for relying on the conclusions of an AI technology that prove to be incorrect, they may use the technology only to confirm their own judgement. While this may shield them from liability, it will discourage use of AI to its fullest potential, which is to augment and not just validate human judgement [254]. If doctors are not penalized for relying on an AI technology, even if its suggestion runs counter to their own clinical judgement, they might be encouraged to make wider use of these technologies to improve patient care or might at least consider their use to challenge their own assumptions and conclusions.

Whether a doctor uses AI also depends on the prevailing standard of care. If AI technologies are viewed as deviating from or are not recognized as meeting the standard of care, doctors will be discouraged from using them, since, otherwise, meeting the standard of care defends (although not absolutely) medical error. If the standard of care requires use of AI technologies, physicians would essentially be mandated to integrate their use into clinical practice [254].

A separate but related issue is the liability of hospitals and health-care systems that select a specific technology. Hospitals could be held liable for failure to exercise due care in selecting the technology or in introducing, using or maintaining it [115]. Generally, a hospital could be held vicariously liable for errors made by clinicians who work at the hospital. Hospitals are thus encouraged both to exercise due care in selecting technologies and to ensure that clinicians have clear guidance on how to use them for both patient care and to avoid errors that result in legal liability for the clinician and the hospital [255]. One possibility would be to establish hospital liability by "negligent credentialing". As, generally, hospitals are liable if they do not adequately review the credentials and practice history of health workers and physicians, they could have a similar duty when introducing AI [256]. For this, hospitals and health systems would have to have the necessary information and tools to identify appropriate AI technologies for clinical use [256]. Hospitals should also have a duty to re-establish control of a process or system that has been automated and that now presents actual or potential risks that were not previously foreseen.

8.2 Are machine-learning algorithms products?

As AI technologies and their software are integrated into or replace medical devices, it is not clear whether they can be characterized as products. Product liability, which holds the manufacturer or developer of a technology or a good to account even if they are not at fault, is a form of strict liability in which liability is imposed even in the absence of negligence, recklessness or intent to harm [257].

Until now, many jurisdictions have hesitated to apply traditional product liability theory to health-care software and algorithms. Product liability could apply insofar as an algorithm is integrated in a medical device or diagnostic. Both European and US courts and new regulations regard medical software as a medical device because of its intended use [258]. Developers may, however, escape liability because in many cases the "actual uses" of a product differ from the "intended uses", even if some of the "actual uses" could have been foreseen [258]. Product liability may also not apply if an AI algorithm is construed as a service and not as a product.

Extension of product liability might be desirable; otherwise, patients might find difficulty in obtaining compensation (e.g., if a clinician followed the standard of care), and bringing a case to assign fault to a developer might be too costly and complex. The design, quality assurance and deployment of AI technologies may involve many people, which could also complicate assignment of liability. Product liability could ensure that developers take all possible steps during development of an algorithm to reduce the likelihood of error, including using diverse, complete data sets to train the algorithm and improving the explainability of the software [259]. Unforeseeable risks and safety failures could, however, limit the effectiveness of current product liability standards.

Assessment of the point to which a developer can be held strictly liable for the performance of an algorithm is complicated by the growing use of neural networks and deep learning in AI technologies, as the algorithms may perform differently over time when they are used in a clinical setting [260] if it is assumed that systems are allowed to update themselves and learn continuously and that use of neural networks and deep learning for AI technologies for health is acceptable and necessary.

Holding a developer accountable for any error might ensure that a patient will be compensated if the error affects them; however, such continuing liability might discourage the use of increasingly sophisticated deep-learning techniques, and AI technology might therefore provide less beneficial observations and recommendations for medical care. It could be argued that liability provisions should be written such as to discourage development of a technology that cannot be fully understood. If this were to be interpreted as requiring the explainability of the mathematical processes that allow an algorithm to learn, however, most machine-learning techniques would be banned. Liability may depend partly on how much control the developer continues to have over an AI technology. In many EU Member States, the extent of a developer's control determines whether a "development risk defence" allows the developer to avoid strict liability [260].

Even if developers could be held strictly liable within a product liability framework, they could avoid liability under the "learned intermediary" doctrine, which limits recovery from a manufacturer when a doctor prescribes drugs or devices [261] for which the manufacturer has provided adequate information, such as warnings about risks [262]. With adequate warnings, decisions by a physician, as the "learned intermediary", break the line of causation between a product developer and the patient who has suffered harm [262].

8.3 Compensation for errors

A liability regime for AI might not be adequate to assign fault, as algorithms are evolving in ways that neither developers nor providers can fully control. In other areas of health care, compensation is occasionally provided without the assignment of fault or liability, such as for medical injuries resulting from adverse effects of vaccines [263]. No-fault, no-liability compensation funds could be supplemented by requiring developers or the companies that develop or fund such technologies to obtain insurance that would pay out for an injury or to pay into an insurance fund, with a separate fund providing compensation when an insurance pay-out is not triggered. In New Zealand, for example, patients seek compensation for medical injuries through a no-fault, no-liability scheme. Injured patients receive Government-funded compensation, thereby giving up the right to seek damages, except in rare cases of reckless conduct [264]. WHO should examine whether no-fault, no-liability compensation funds are an appropriate mechanism for providing payments to individuals who suffer medical injuries due to the use of AI technologies, including how to mobilize resources to pay any claims.

8.4 Role of regulatory agencies and pre-emption

AI technologies, like drugs and devices, will be increasingly subject to regulatory oversight and validation before use, especially as their uses expand and as clinicians increasingly rely upon them. If a commercial algorithm is approved by a regulatory agency, the doctrine of pre-emption may apply, i.e., that a decision taken by a central government agency to validate a technology will supersede any cause of action guided by civil laws [265]. Pre-emption may not always be relevant, however, especially if the regulatory pathway for approval of an AI technology is abbreviated or regulatory approval is based on little information on how the algorithm was constructed and trained and may perform over time [265]. Furthermore, as developers in some jurisdictions may not be held accountable for an algorithm as it evolves and learns after its sale, a doctrine of pre-emption may not be applicable if an algorithm evolves after a regulatory agency has approved the technology.

8.5 Considerations for low- and middle-income countries

Much of the literature, policy frameworks and court decisions on liability regimes are from the EU and the USA, which is where AI technologies are actively deployed. It is not known whether these approaches will be adopted in LMIC or whether those countries will take different approaches to liability. Liability rules play an important role in promoting safety and accountability, and, in some cases, they are the first and only line of defence against errors made by machine-learning technologies. Many LMIC still lack sufficient regulatory capacity to assess drugs, vaccines and devices and might be unable to accurately assess and regulate the rapidly arriving machine-learning technologies for the public good. Concern that such technologies might not operate as intended is heightened by the lack of good-quality data to train algorithms and the fact that AI technologies may have "contextual bias" [192]. Such concern should not preclude the use of AI in LMIC, but it highlights the importance of robust, effective liability regimes. Many LMIC may wish to use AI technologies in resource-poor settings for reasons that do not apply in the EU or the USA, such as lack of health-system infrastructure.

In many LMIC, injured parties may not have access to justice, or it may be too expensive or too protracted, so that it not just difficult to obtain compensation for harm caused by AI technologies but it is also unlikely to serve as a deterrent to those responsible for the development and deployment of

such technologies. Marginalized populations have even less protection and are often excluded from redress within the legal system. It might also be difficult to seek compensation if the AI technology was developed by an international company or developer with no physical presence where the harm occurs. These challenges must be addressed to increase the effectiveness of liability rules.

LMIC might have to address challenges and risks that are not often considered in high-income economies. These include lack of appropriate training data for the algorithm to ensure that it performs accurately for patients with a different physical appearance and poor connectivity, which can compromise reliable, safe use of a technology.

Even if legal systems in LMIC adopt the approaches of HIC for the introduction of AI technologies for clinical use, they will have to develop approaches that are consistent with legal practices and standards to compensate people who are harmed by such technologies, hold companies and governments accountable for the products they develop and calculate the risk-benefit for using or refusing AI technologies. WHO should work with other United Nations agencies and with governments in the design and introduction of appropriate liability rules.

Recommendations

1. International agencies (and professional societies) should ensure that their clinical guidelines keep pace with the rapid introduction of AI technologies, accounting for the evolution of AI technologies by continuous learning.
2. WHO should support national regulatory agencies in assessing AI technologies for health.
3. WHO should support countries in evaluating the liability regimes that have been introduced for the use of AI technologies for health and how such regimes should be adapted to different health-care systems and country contexts.
4. WHO and partner agencies should seek to establish international norms and legal standards to ensure national accountability to protect patients from medical errors.

9 Elements of a framework for governance of artificial intelligence for health

Human rights standards, data protection laws and ethical principles are all necessary to guide, regulate and manage the use of AI for health by developers, governments, providers and patients. Many stakeholders have called for a commonly accepted set of ethical principles for AI for health, and WHO hopes that the principles suggested in this document (See section 5.) will encourage consensus.

Use of AI for health introduces several challenges that cannot be resolved by ethical principles and existing laws and policies, in particular because the risks and opportunities of the use of AI are not yet well understood or will change over time. Furthermore, many principles, laws and standards were devised by and for HIC. LMIC will face additional challenges to introducing new AI technologies, which will require not only awareness of and adherence to ethical principles but also appropriate governance.

Governance in health covers a range of steering and rule-making functions of governments and other decision-makers, including international health agencies, for the achievement of national health policy objectives conducive to universal health coverage. Governance is also a political process that involves balancing competing influences and demands.

At the Seventy-first World Health Assembly in 2018, Member States unanimously adopted resolution WHA71.7, which calls on WHO to prepare a global strategy on digital health to support national health systems in achieving universal health coverage [266]. A global strategy and other governance frameworks and standards established by WHO will contribute to a governance framework for AI for health. This section addresses the ethical dimensions of several areas of governance.

9.1 Governance of data

The definition of "health data" has widened dramatically over the past two decades. Successful development of an AI system for use in health care relies on high-quality data, which are used to both train and validate the algorithmic model. This section addresses the evolution of individual consent with the proliferation of health data as well as the principles, legal frameworks and measures used by governments. This section also addresses principles and mechanisms designed and used to govern health data by communities, academic or health-care institutions, companies or governments, including how these entities should share health data.

9.1.1 Evolving approaches to consent

As the types, quantity and applications of health data, including for commercial use, have grown, a patchwork of approaches has emerged to facilitate individuals' relation to their health data. The main challenge is safeguarding individual privacy and autonomy by controlling their data without limiting the purported benefits of their collection and use. These considerations are likely to apply whether the data are used for AI or for a relational database.

Mechanisms for individual control of data, such as informed consent, a duty of confidentiality and de-identification, may not be sufficient and may interfere with positive uses. (See section 6.3.) Therefore, several "modified" approaches to consent could be used as the quantity of health data and their possible uses increase. Consent must be given only after explanation of the consequences of providing it, including for example which data will be used and how and the consequences if consent is not given.

One form of consent that could improve individual control and choice is electronic informed consent, in which online forms and communication are used to give consent for various uses of health data [114]. Electronic informed consent could allow users better understanding of how their data will be used and improve their control of the data. The content should, however, be presented simply so that it is readily accessible to the general public, such as with illustrations, to ensure that consent is given freely and that the risks are understood [114]. Sage Bionetworks, for example, has established a toolkit and information guide for facilitating provision of electronic informed consent [267]. Another approach is "dynamic consent", which allows users to modify their consent periodically for uses that they wish to permit and those that they specifically exclude [114]. A third approach to consent, discussed below, is to seek "broad consent" from individuals to facilitate secondary use of health data without undermining their rights to privacy and autonomy.

Alternatively, governments might wish to define when consent can be waived in the public interest. This is already permissible under data protection laws if it is strictly necessary and proportionate to achievement of a legitimate aim. This implies that, in certain situations, government could have a duty to share health data for the benefit of the wider public or for other non-monetary benefits, such as better quality of life or health [268]. Thus, consent would be waived because the data are considered a public good for which data can be "conscripted for publicly minded uses" [128]. This could include situations in which there are clear public health benefits of using data that would otherwise be unavailable because too many individuals have opted out of sharing such data. The burden of demonstrating that lack of consent is undermining a benefit should rest with the entity that seeks to avoid consent. It could imply that obtaining health data without the specific consent of the individual is justified if the benefit is broadly distributed and outweighs violation of privacy when the risk is "low" [128]. A system in which benefits and risks are weighed could, however, invariably lead to sharing of data without consent, as medical benefits – whether better surveillance of disease or development of a new drug – could always be considered more important than a "low risk" of violation of privacy from use of the data.

Another concern is that a government or a company may define "public interest" in a way that is not based on public health or patient need. Whether patients share the benefits may depend on the entity with which they are shared, such as commercial actors, which may not share benefits if the medical

products and services are neither affordable nor available (see below). Thus, conscripting health data with the broad goal of contributing to the public good is questionable when the data are shared with a commercial entity, whatever the intended product or service. Recent instances (described in Section 6.3) of patient data that were shared by not-for-profit entities or academic institutions with private companies without the consent of the patients has raised significant concern, as the patients were not notified that their data were shared, for what purpose or the identity of the private entity.

In Japan, an approach to resolving such conflicts was passage of the Jisedan Iryo-kiban Ho (Next Generation Medical Infrastructure Law), which permits hospitals and clinics to provide patient data to accredited private sector companies, which are responsible for making the data anonymous and searchable [269]. Before sharing data, hospitals and clinics must inform patients and give them the right to opt out. The accredited data companies anonymize and store the data and make it available to academic researchers, pharmaceutical companies and government agencies for a fee. Accredited data companies are required to institute safeguards for cybersecurity, unauthorized use of data and unauthorized disclosure by employees [269].

In 2020, the EU proposed a means for use of data without consent under the concept of "data altruism", previously known as "data solidarity" [270]. This would allow companies to collect personal and non-personal data on individuals for projects that are in the public interest. The approach seeks to limit the type of company that can collect data by specifying that it must: be constituted to meet objectives of "general interest"; operate on a not-for-profit basis and be independent of any for-profit entity; ensure that any activities related to data altruism are undertaken through a legally independent structure separate from its other functions; and can voluntarily register as a "data altruism organization" in an EU Member State. To facilitate data altruism, a common European consent form will be developed, which can be tailored for different sectors and uses.

Data altruism could raise concern. First, this form of data-sharing could lead to exceptions or "grey areas" in which health data are used for commercial purposes for which the individuals from whom the data were obtained would not wish to provide consent. Secondly, such a regulation could be rewritten over time to redefine the entities allowed to collect data for altruistic purposes. Thirdly, even if the health data were initially used for a non-commercial objective, such as in drug discovery, the product or service that emerges might eventually be licensed to or acquired by a commercial entity rather than remaining in the public domain.

9.1.2 Broad consent

Several not-for-profit institutions that have deposited health data in centralized biorepositories practise principles of informed consent for sharing such data, which ensures that the person who provides data understands consent at enrolment. Any industry partner is disclosed at the time of consent, and prospective, explicit consent is given for future secondary use of the data for research [271]. These standards do not prevent secondary use of health data, except when, for example, commercial actors that were not included in the initial consent seek to use the data or when commercial actors could otherwise gain access because they subsidize activities of not-for-profit entities that have access to the data. Even with additional standards in place, at a biorepository operated by the University of Michigan, USA, access to data was denied by a review committee for only 6 of 70 projects proposed over 2 years and only because of inadequate initial consent [271].

Another concern with use of health data for research arises when the data are user-generated, such as data obtained from digital devices and wearables and data supplied by users to social media and other platforms and to online patient communities. Governance of such data, which may not have been collected initially for research, is complex because of the "lack of international boundaries when using the internet" and because the "online information industry has failed to self-regulate" [133]. Andanda suggested that one means for improving governance of such data would be to encourage health researchers to adhere voluntarily to the "Global Code of Conduct", which encourages researchers and institutions to develop context-specific codes, be fair, respectful, caring and honest when dealing with online users and practise ethically informed research practices [133].

A more controversial issue is creating a market or system through which individuals can buy and sell health data. Health data are sensitive personal data, linked to human agency and dignity. A system that facilitates the sale of personal data could lead to a two-tier society in which the wealthy can protect their rights and afford to limit use of their data by other parties, whereas people living in poverty may feel compelled to sell their data to access social or material benefits. A system that facilitates the sale of data would be in contravention of several human rights standards. Furthermore, while the sale of data might contribute to uses that are commercially valuable but less beneficial to individual or public health, the data market itself may not function properly and could undervalue an individual's data. The sale of data could lead to loss of control by an individual of his or her health data. Such challenges with health data have emerged with commercial sale of blood and related products such as plasma [272].

9.1.3 Data protection

From a human rights perspective, an individual should always control his or her personal data. Individuals' right to their own data is grounded in concepts that are related to but distinct from ownership, including control, agency, privacy, autonomy and human dignity. Control may include various approaches to individual consent (see above) and also collective mechanisms to ensure that the data are used appropriately by third parties (see below). Data protection laws are rights-based approaches that include standards for the regulation of data-processing activities that both protect the rights of individuals and establish obligations for data controllers and processors, both private and public, and also include sanctions and remedies in case of actions that violate statutory rights. Data protection laws can also provide for exceptions for non-commercial uses by third parties. Over 100 countries have adopted data protection laws [273].

Data protection frameworks and regulations are essential for managing the use of health data. The EU GDPR, which applies to citizens and residents of the EU, irrespective of whether the data controller or processor is based in the EU, also has a global reach because it applies to non-EU citizens or residents if the data controller or processor is based in the EU. The GDPR is designed to limit the data collected about an individual to only that which is necessary, to allow collection of data only for listed legitimate purposes or with an individual's consent, and to notify individuals of data-processing activities. Health data are protected under GDPR unless an individual provides specific consent or if use of the data meets certain exceptions, such as for health-related operations or scientific research. Even when exceptions apply, data processors and controllers must respect certain obligations.

GDPR also introduced "data portability", the right of individuals to obtain their personal data in a machine-readable format from one controller that can be sent to another controller [113]. Depending on how data portability is implemented in the EU, it could allow individuals to control their own data and to share them with additional entities. Data portability could decentralize the control and distribution of data and, with appropriate implementation, could be a novel form of data management that fosters both oversight and innovation.

Data protection regulations are enforced by data protection authorities, which develop and administer regulations, provide guidance and technical advice and conduct investigations. South Africa, which introduced a data protection regime for the first time in July 2020 with enactment of the Protection of Personal Information Act 4, will introduce enforcement in mid-2021 through several means, including administrative fines that could exceed US$ 500 000 and also civil cases and criminal liability [274].

Some governments have nominated additional supervisory authorities to facilitate the use of health data. The United Kingdom established a National Data Guardian in 2014 for appropriate management of health data with respect to confidentiality and to improve the use of such data for beneficial purposes. In 2018, the entity was granted the power to issue official guidance on the use of data for health and adult and social care in England [275].

9.1.4 Community control of health data – data sovereignty and data cooperatives

Measures have been taken not only to promote the individual right to privacy and autonomy over health data but also to provide discrete communities with control over their data, including health data, through the exercise of data sovereignty or creation of data cooperatives. Several indigenous communities have sought to establish control over their data through data sovereignty. Māori (the indigenous population of New Zealand) have introduced principles for data sovereignty that establish, for example, control over data, including to protect against future harm, accountability to the people who provide such data by those who collect, use and disseminate them, an obligation for such data to provide a collective benefit, and free prior and informed consent, which, when not obtainable, should be accompanied by stronger governance [276]. Māori also recognize that the individual rights of data holders should be balanced by benefits for the community and that in some situations the collective rights of the Māori will prevail over those of individuals [276].

First Nations groups in Canada have also outlined principles for sovereignty over their data, with four elements: ownership of data, control of data, access to data and possession of data. It is expected that, over time, First Nation tribes will establish protocols to allow wider access to these data for uses that benefit them [277].

A data cooperative gives people who provide data control over their data by storing the data for the members of a cooperative. Data cooperatives allow secondary uses of such data while allowing members of the cooperative to decide collectively how the data should be used [113]. Data cooperatives allow members to set common ethical standards, and some have developed their own tools and applications to ensure that the data are used beneficially [113].

9.1.5 Federated data

Federated data systems have grown significantly. They include collaborations between research institutions, governments and the public and private sector and within the private sector. Federated data-sharing has been defined as "a promising way to enable access to health data, including genomic data, that must remain inside a country or institution because of their sensitivity" [278]. Data do not leave the participating organization that holds them, but authorized users can make queries that allow them to access data, for example to train an algorithm. Proponents have noted that federated data systems allow each entity to govern use of its data and that the approach preserves privacy and security [278]. While federated data-sharing may facilitate analysis of large data sets while maintaining local control, it does not overcome concern that informed consent might not have been sought for secondary uses of the data [137].

9.1.6 Government principles and guidelines

Some governments that are collecting and using health data for commercial and public sector interventions have established principles for data collection and use. The United Kingdom's NHS has established five guiding principles for a framework in which data can be used in health innovation. A notable commitment under these principles is transparency – that any commercial arrangements should be transparent, clearly communicated and not undermine public trust or confidence [279]. As discussed below, however, many agreements between the public and the private sector are not transparent, which raises serious concern if there are also financial conflicts of interest.

Other forms of transparency could be required, such as the transparency of sources and methods of obtaining and processing data, how and why certain types of data are excluded, the methods used to analyse the data and open discussion in publications of data bias.

In New Zealand, an independent ministerial advisory group funded and appointed by the Government conducted a wide-ranging consultation to build an "inclusive, high-trust, and high-control data-sharing ecosystem" [280]. The guidelines include eight questions about what matters most to people in building trust in data use and whether the use of data provides value, protection and choice for an individual (Figure 2).

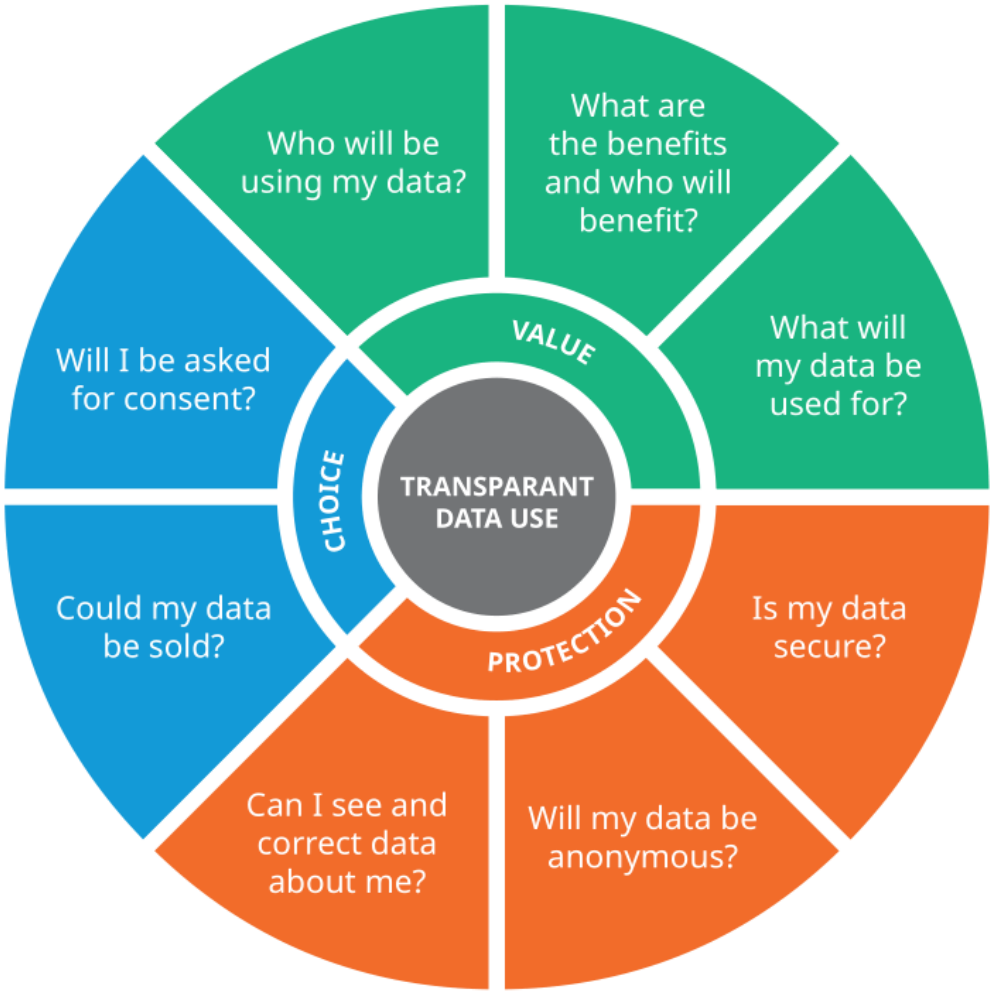


Figure 2 – Elements of transparent data use [280]

Although the guidelines are voluntary, each entity that seeks to use the data has been asked to publish answers to these questions so that the individuals who provide the data can determine whether the values of the entity align with their preferences [280].

WHO has introduced its own data principles [281], which are designed to provide a framework for data governance by WHO and to be used by staff to define the values and standards that govern how data that flow into, across and out of WHO are collected, processed, shared and used. The five principles are as follows.

1. WHO shall treat data as a public good.
2. WHO shall uphold Member States' trust in data.
3. WHO shall support Member States' data and health information systems capacity.
4. WHO shall be a responsible data manager and steward.
5. WHO shall strive to fill public health data gaps.

WHO is also introducing a data governance framework that would introduce the necessary standards, solutions and structures to ensure the quality and integrity of WHO data, from collection, storage, analysis and validation through to use. To ensure that the principles can be put into practice, WHO will use a "hub-and-spoke" governance model to obtain feedback and approval, and data focal points at WHO will work with regional focal points on issues that arise during the ever-growing use of health data. They will also be guided by the Data Governance Committee constituted by WHO [282].

9.1.7 Data-sharing, including data hubs

As health data have proliferated, governments have taken steps to improve data-sharing for scientific research and also for commercial development of health AI and other health applications. In 2014, the US National Institutes of Health introduced their Genomic Data Sharing Policy, which is intended to encourage "broad and responsible sharing of genomic research data" [283]. Legislation enacted in the USA in 2016, the 21st Century Cures Act, extended the remit and created statutory authority of the Director of the National Institutes of Health to require researchers who received awards from the Institutes to share their data and to provide the means for the Institutes to enforce data-sharing [284].

The Act also provides means to improve the access of individuals to their own health data, which was finalized in rules issued by the US Government in 2020 that create a requirement for health information technology providers to introduce a standards-based application programming interface to support an individual's use and control of electronic health information [285]. Health information technology providers must meet three requirements for its interface to be certified: it must meet certain technical programming standards that ensure interoperability, it must be transparent, and it must be "pro-competitive" or promote efficient exchange, access and use of health data [285]. The requirements for health information technology providers, such as anti-blocking or interoperability, show that governments can mandate and manage commercial use of AI and other technologies for health care.

9.1.8 Data hubs

Numerous data hubs pool various types of health data for use by third parties, which depend on the type of data hub. Several government-sponsored data hubs have emerged. In the USA, two such hubs are the Precision Medicine Initiative (All of Us) [286] and the Department of Veteran Affairs health data hub. The EU is establishing a European Health Data Space to facilitate the exchange and sharing of health data (e.g., health records, genomics, registries) for purposes such as the delivery of primary care and the development of new treatments, medicines, medical devices and services, while ensuring that people have control of their own health data [287].

Health Data Research UK is an independent, not-for-profit organization of 22 research institutions in the United Kingdom that collect health data and make it available to public and private entities for research on diseases and ways to prevent, treat and cure them. Principles of participation have been defined in consultation with policy-makers, the NHS, industry and the public [288].

9.1.9 Data-sharing and data partnerships with the private sector

One of the more difficult questions in the creation of government, not-for-profit or academic data hubs is how they should work with companies, either in accepting data that could improve their quality or allowing the companies to use their data for training or validation of algorithms. When commercial entities make use of such data, there is concern, which has sometimes materialized, that the people from whom they were derived did not knowingly given consent for their use for commercial purposes. There is an additional concern that such agreements are not disclosed to the public or to private sector parties to such agreements.

For example, numerous agreements signed between the Mayo Clinic, a major health system in the USA, with 16 technology companies provided the Clinic with a "revenue stream and generated crucial insights for health tech firms eager to commercialise digital products and services" [137]. In some cases, the Clinic not only shared data with a company but subsequently took an equity stake in those companies, which provided the Clinic with additional revenue. De-identified patient data were shared without requesting consent or even notifying the people who had supplied their health data for products under development. The names of eight of the firms that signed agreements were not disclosed, and none of the contracts signed between the Mayo Clinic and its technology partners were made public [137].

In other cases, physicians or scientists in health-care systems who had access to raw data provided to health technology firms founded or invested in the companies. An investigation in 2018 found that board members and senior executives at the Memorial Sloan Kettering Hospital in the USA had either founded or invested in an AI start-up to improve cancer diagnosis and had used the Hospital's trove of 25 million patient tissue slides and six decades of pathology research for the company's benefit without open bidding or transparent consideration of whether the data should be shared. Memorial Sloan Kettering had also taken an ownership stake in the company [289].

Some companies, either alone or in collaboration with other companies, have established health data hubs with data from one or more companies, which are used in the development of products and services. Such partnerships, which may result in useful products and services, raise concern about the transparency of the activities, oversight of activities, competition and whether such private carriers of data will seek consent or at least engage the communities and individuals that provided the data.

Recommendations

1. Governments should have clear data protection laws and regulations for the use of health data and protecting individual rights, including the right to meaningful informed consent.
2. Governments should establish independent data protection authorities with adequate power and resources to monitor and enforce the rules and regulations in data protection laws.
3. Governments should require entities that seek to use health data to be transparent about the scope of the intended use of the data.
4. Mechanisms for community oversight of data should be supported. These include data collectives and establishment of data sovereignty by indigenous communities and other marginalized groups.
5. Data hubs should meet the highest standards of informed consent if their data might be used by the private or public sector, should be transparent in their agreements with companies and should ensure that the outcomes of data collaboration provide the widest possible public benefit.

9.2 Control and benefit-sharing

The application of big data and AI for health care raises questions about how to assess and govern data control, IP and other proprietary and privacy rights that might affect the use and control of medical data and AI-driven technologies. These include asserting exclusive rights over health datasets, algorithms, software and products that include AI and the outcomes of AI-based technologies, such as medicines and diagnostic technologies. Several wider questions should be resolved, including whether health big data can or should be controlled exclusively by individuals by an appropriate form of governance or by entities that may aggregate the data. (Control of personal data is discussed above.)

A separate question is whether novel products created solely by a machine can be "owned" and, if so, whether ownership rights are conferred on the machine or on the entity that created or controls the machine. There is also the question of assigning appropriate value to the public's contribution to development of new AI technologies, such as investment in the development of algorithms, provision of data by individuals and health systems and from health data hubs accessed by private actors for the development of new AI technologies. If AI technologies are increasingly protected by exclusive rights, there is the wider question of whether they will be available, appropriate and affordable in LMIC.

9.2.1 Control over and benefit-sharing of big data

The central role of big data for AI, including medical big data for use of AI for health care, has led to labelling of data as the new "oil", a valuable commodity over which there will be increased commercial conflict for its control, use and access [290]. Such labelling has been criticized as

unhelpful and conceptually inaccurate ([291], [292]). Unlike oil, the supply of data is virtually infinite, and they can be re-used in other contexts with valuable commercial or non-commercial applications. There is at least the possibility of control of and consent for use of one's data. While the intrinsic value of oil is captured once it is extracted or drilled (subject to processing and refining), data are not intrinsically valuable unless data science is used to generate something of value.

Another view is that it is not so much the commercial value of data but its use in the development and deployment of AI-based applications that is important. In this view, data are the "oxygen", an indispensable resource for the public infrastructure required for AI and data science to serve the public and private sectors [293]. Whether data should be considered "oil" or "oxygen" (or neither) depends partly on whether exclusive rights can or should be associated with data, who should have such exclusive rights and to what extent they should impede others from access to and use of the data for public or private uses.

Several types of IP rights may apply to data and software, including protection of trade secrets, copyright, database rights (in only a few jurisdictions), regulatory exclusivity and, in rare circumstances, patent rights. Data and software as such cannot be patented in most jurisdictions, but "functional" data used in technical applications may be patented ([294], [295]). It is beyond the scope of this publication to discuss the IP rights that could apply to large data sets or to big data, yet such rights, if they are to be expanded or minimized with respect to large data sets or big data depend on broader policy objectives and ethical considerations.

There is a conflict between sharing data and the commercial prerogatives that are protected by IP rights [296]. On the one hand, conferring IP and related rights to health big data could discourage open sharing of the data, which is necessary to advance scientific progress and the development of AI for health care and medicine ([93], [295]). Public or private "owners" of health big data might not grant third parties the right to use the data to develop novel AI technologies, thereby undermining open innovation [297] and giving commercial entities the power to exclude competitors or engage in "rent-seeking". Questions should arise about who is allowed access, the rationale for inclusion or exclusion and the conditions under which the data will be accessible (including whether fees must be paid), especially for third parties that wish to use the data for non-commercial purposes. On the other hand, lack of IP rights to health big data could discourage some commercial investments [297]. While the 21st Century Cures Act, enacted in the USA in 2016, encourages the sharing of data (see section 9.1), it asserts that proprietary interests supersede data-sharing interests and that the ability of the US Government to mandate data-sharing is limited by policies for prioritizing the protection of trade secrets, proprietary interests, confidential commercial information and IP rights [284]. Similar consideration apply, for example, to the FAIR Data principles of the European Open Science Cloud, which plans to create data-sharing clouds that are "as open as possible and as closed as necessary" and does not preclude respect for IP rights or the protection of privacy rights [298].

An additional concern is whether sharing of health data by communities, health systems or governments in LMIC will include sharing of benefits, especially if the data are used for commercial applications of AI [93]. If benefits are not shared, it may be either because there are no legal conventions or frameworks that mandate benefit-sharing of the uses of big data or because the entities that negotiate benefit-sharing on behalf of LMIC may have to negotiate from a weaker position [295]. Benefit-sharing may include not only equitable access to and availability of technologies that arise from sharing health big data but also the assurance that enough investment is made in digital infrastructure, research capacity, training and infrastructure to ensure that the products of AI and big data are also generated by researchers and companies in LMIC [295]. New technologies that require "state-of-the-art" capacity, such as quantum computing, might exacerbate inadequate benefit-sharing.

Thus, while IP rights could be adjusted case by case to encourage open innovation, investment or benefit-sharing, control (and IP rights to assign control) may be inappropriate to encourage widespread use and application of health data, in view of numerous competing considerations, including an individual's right to privacy and control [299], society's interest in scientific progress

and the development of AI-guided technologies, commercial interest in exploiting such data for profitable activities and the interest of data contributors (communities, health systems, governments) in sharing the benefits generated by third parties [299].

It has been recommended that the focus be not on recalibrating or introducing new IP rights, which could impede data-sharing or intensify competing claims to control of data, but instead on establishing a legal framework based on custodianship [93]. Custodianship, or responsible oversight with ethical values, can ensure access to data, promote fair data-sharing and preserve privacy. While those who provide data maintain limited control, certain decisions are delegated to data custodians with custodial rights – and not control (or IP rights) – over big data. Custodial rights can include protecting the privacy of those who contribute data, disseminating research findings, ensuring freedom of scientific enquiry and providing attribution to those who invest in creating databases and agreeing on terms of use and access [295].

9.2.2 Ownership of AI-based products, services and methods

Products and services created with AI and big data could be patented or subject to other IP rights. These include algorithmic models that can be used in drug discovery and development and the end-products of such uses of AI, such as new medicines, medical devices or diagnostic methods. Thus, as noted in section 3.2, the announcement by DeepMind of a new AI model, AlphaFold, may result in real progress in the development of new medicines but might be heavily protected by patents and other forms of IP and therefore not widely available. If other AI technologies and tools that could accelerate drug development are not placed in the public domain (e.g., without IP protection) and are not available for licensing on a royalty-free basis or under reasonable terms and conditions, the companies that own such technologies will exert greater power and control over the development of new medical technologies and services.

An overlying concern in patenting (and other forms of ownership) of AI-generated inventions is therefore that IP rights could exclude affordable access to the products or services and that patent holders engage in rent-seeking behaviour to recuperate investments and earn outsized profits. As novel medicines, diagnostic methods and other products and services developed with AI may depend on publicly generated health data and other public-sector investments in AI and health-care infrastructure for identification, testing and validation, the question arises of whether the public investment will be rewarded, including by ensuring affordable access to the product. All science, including advances in AI, has been based on decades of publicly funded academic research.

Assessing ownership is especially difficult when a product or research output is the result of a PPP for which governments may have provided funding and other forms of support but which maintain limited or no ownership of the research output. Ensuring a role for government in both the development of new AI technologies and the ownership of the outcomes would be fairer for the governments and citizens that contribute resources and data to collaboration with the private sector.

Another concern is that issuing time-limited patent monopolies for such inventions, even if they encourage innovation, may discourage the companies that own AI technologies from considering the needs of people living in poverty in LMIC when developing or adapting such products. Thus, as AI is used more frequently to develop new technologies to improve health care, including new medicines, the use of incentives outside the patent system, such as those that separate the cost of research and development from the expectation of high prices, could encourage companies that develop these technologies to invest in use of AI or to adapt new products to meet global public health needs.

Companies might refuse to disclose data that they consider an "essential facility" for developing, for example, a much-needed vaccine or choose to collaborate only in strategic areas of data application and with control of the data that are shared, with whom and under which conditions. This could replace healthy competition by collusion, with future effects on competition that are difficult to

assess. Antitrust (competition) authorities will have to consider new approaches to address such issues [297].

Several legal issues will affect the patenting of AI technologies. One is whether AI-guided machines that develop new products or services can be considered inventors, which would lead to questions about defining the threshold for meeting the criteria for patenting an invention, such as an inventive step. Some legal experts have argued that recognition of machines as inventors would encourage the development of creative, powerful machines that can generate new innovations [300]. If, however, most such machines are owned by a few companies, the benefits of the inventions will accrue to those few companies, which will wield significant power through exclusive rights and use the machines to capture an entire field of technology. In January 2020, the European Patent Office ruled that machines cannot be listed as inventors under current patent laws [301], and the US Patent and Trademark Office has issued a similar decision [302].

Another legal issue is whether diagnostic methods and algorithms can be patented. While in the USA securing patent protection for diagnostic methods and mathematical models is highly restricted, the EU has provided several grounds for the issuance of patents [303]. While patent monopolies could encourage the development of new technologies with greater medical benefits, patenting of such methods and services could limit their diffusion, access and benefit-sharing with the populations that contributed the data used to train or validate the technology.

Recommendations

1. WHO should ensure clear understanding of which types of rights will apply to the use of health data and the ownership, control, sharing and use of algorithms and AI technologies for health.
2. Governments, research institutions and universities involved in the development of AI technologies should maintain an ownership interest in the outcomes so that the benefits are shared and are widely available and accessible, particularly to populations that contributed their data for AI development.
3. Governments should consider alternative "push-and-pull" incentives instead of IP rights, such as prizes or end-to-end push funding, to stimulate appropriate research and development.
4. Transparency in regulatory procedures and in interoperability should be enhanced and should be fostered by governments as deemed appropriate.

9.3 Governance of the private sector

The private sector plays a central role in the development and delivery of AI for health care. The "private sector" ranges from small start-ups to the world's largest technology companies, as well as companies that provide many of the materials necessary for AI, including health data collected by companies that supply wearable devices, data aggregators and software firms that write new algorithms for use in health care. Furthermore, many companies that were already providing products and services are transforming their businesses to integrate AI and big data. These include biopharmaceutical companies, diagnostic and medical device firms, insurance companies, private hospitals and health-care providers. Companies that are developing AI technologies for use in health care are also providing these applications and services outside the health-care system, raising the question of how such health-care provision should be regulated.

This section addresses several issues related to the governance of such companies: To what extent should oversight and governance of the private sector be enforced by companies collectively or individually? What challenges and opportunities for effective governance are associated with PPPs for AI for health care? What are the challenges of oversight and governance of large technology companies involved in the use of AI for health? How should governments manage the growth of

health-care services provided by companies outside the health system? How can governments ensure that they are effectively overseeing the private sector?

9.3.1 The role of self-governance

As companies often push the boundaries of innovation and act much more quickly than can be anticipated by regulators, governments and civil society, they often first set the rules in the code that they write, the services they design and the corporate practices and terms of services they offer [304]. As some innovations have raised concern, companies have strengthened their internal processes and measures to avoid criticism and have pursued collaborations and partnerships. Thus, some have introduced their own ethical principles and internal processes for integrating ethical considerations into their business operations [156]. This includes integrating ethics into the design of new technologies and design-related approaches to privacy and safety. Companies have also launched multi-stakeholder initiatives to develop best practices [305], although there is no such initiative yet for the use of AI for health.

While integration of ethics into a company's operations is welcome, it raises as many concerns as hopes, the concerns including that companies may be engaging in "ethics-washing" and that the measures are intended to forestall regulation instead of adapting to oversight [156]. In some companies, efforts by ethics teams to address ethical challenges and concerns may be discouraged or have repercussions. For example, a news report stated that Google had fired an AI ethics researcher who criticized Google's "approach to minority hiring and the biases built into today's artificial intelligence systems" [306]. Even if attempts to formulate and integrate ethics into daily company operations are taken seriously, other challenges may limit their effectiveness.

First, the incentives and values of AI firms and developers may differ from those of the patients, health-care providers and health-care systems [306] that will use such products and services but have no role in establishing the culture or norms in which the products and services are developed [307]. For example, large technology companies, which are based in only a few countries, may adopt values and belief systems that are not appropriate for other countries, health-care systems or communities. More generally, while medicine is guided by the objective of promoting the health and well-being of patients, an AI developer who is developing a product or service that provides benefits is ultimately working in the interests of the company to develop a profitable service or product and, in the case of publicly traded companies, for their shareholders [305]. While medical professionals have a long-standing fiduciary relationship with patients, AI developers, however well-intentioned and with emerging expectations and legal obligations to protect individual privacy, have no fiduciary duty to patients or health-care providers. This complicates any attempt by an individual or a company to put the health and well-being of patients first [305].

Secondly, the ethical norms adopted by companies might be difficult to translate into practice [156], either because AI developers have no suitable methods of doing so, as AI is a relatively new technology, or practical measures to adhere to high-level ethical norms may be difficult to reconcile with a culture of fast growth, fast failures and getting first to the market. Ethical principles may therefore be "watered down", modified or rendered ineffective. It may also be difficult to determine whether ethical norms are written into the source code for an AI technology, whereas, in the practice of medicine, numerous structures built over time, including professional societies and boards, ethics review committees, accreditation and licensing schemes, peer self-governance and codes of conduct, determine and shape what is acceptable, and bad practices and bad actors can be identified quickly [305].

Thirdly, there are insufficient legal and professional accountability mechanisms to reinforce good-faith efforts of firms to turn ethical principles into practice [305]. Unlike the medical profession, AI developers and technology firms have no effective self-governance mechanisms and do not face the legal penalties and repercussions of other professions, especially the medical profession. Accountability mechanisms in the medical profession reinforce its fiduciary duty to patients and are

reinforced by sanctions to deter poor practices. AI development does not include professional or legally endorsed accountability mechanisms [305].

Fourthly, it is questionable whether companies can govern their own AI products and services effectively to minimize any harmful direct or indirect impact on health care. For example, social media companies such as Facebook play an important role in sharing health information through platforms such as Facebook and WhatsApp. There has recently been significant concern about the spread of misinformation and disinformation on its platforms that undermines medical and public health information issued by governments and international agencies, and this has increased during the COVID-19 pandemic. The company has taken steps to address misinformation and disinformation, including a partnership with WHO to create a chatbot on Facebook Messenger and WhatsApp to provide accurate information through the WHO Global Alert Platform [308].

A study by a not-for-profit group, Avaaz, found, however, that the spread of medical disinformation and misinformation on Facebook far exceeded information from trustworthy sources such as WHO. The most popular "super spreader" sites received four times more clicks than bodies such as WHO and the US Centers for Disease Control and Prevention [309]. According to Avaaz, this was due largely to amplification of public pages that featured misinformation in Facebook's algorithm. During the early stages of the COVID-19 pandemic, in April 2020, "disinformation sites attracted an estimated 420 million clicks to pages peddling harmful information – such as supposed cures for SARS-CoV2" [310]. Only 16% of misleading or false articles displayed a warning label by Facebook third-party fact-checkers [310]. Furthermore, while Facebook has subsequently sought to address misinformation on COVID-19 by deleting false posts and directing users to valid information [311], some researchers have criticized Facebook for not identifying the misinformation and correcting it [312].

The concern that a few companies manage information critical to the public good extends to whether such companies might withhold such information because of public policy or corporate disputes. In 2021, Facebook, having been unable to reach an agreement with the Australian Government about a new law that would require the company to pay news publishers for the content it placed on its site, decided to block users from accessing news stories on its platform [313]. The block included access to Australian state government health websites and prevented the state governments from posting on the website, even as the Government was preparing public announcements about vaccination against COVID-19 [314]. Websites that posted misinformation about vaccines were unaffected [315].

None of these concerns should be a reason for companies not to invest in improving the design, oversight and self-regulation of their products. The improvements could include licensing requirements for developers of "high-risk" AI, such as that used in health care, which would bring AI developers in line with requirements in the medical profession and increase trust in their products and services. International standards organizations have made important contributions to improving applications of health information technology, from data structure and syntax to privacy and implementation. For instance, the International Standardization Organization [316], Health Level Seven International [317] and other organizations have contributed to the governance of information technology, including machine learning, and such standards have been described as carrying ethical weight [177].

9.3.2 Public-private partnerships for AI for health care

PPPs are common in health care, and, unsurprisingly, PPPs are emerging in the field of AI for health care. In one type of PPP, raw data are provided by the public sector, such as electronic medical records and other health data collected in health-care systems and hospitals, and these are used by one or more companies to develop products and services, such as diagnostic methods and predictive algorithms.

Supporters of PPPs in both government and industry emphasize the benefit of leveraging the resources and innovative capacity of companies to generate products and services. Presumably, in

such collaborations, governments can oversee the activities of the private companies and safeguard the public interest. There are, however, challenges in ensuring effective governance of the private sector. First, there is a significant asymmetry in information and skills between companies and government agencies in such partnerships. Companies often hire trained professionals who are well versed in the technology in question and in the parameters of a negotiated partnership. A second challenge is that the "social license" granted to the public sector for use of certain resources, such as patient data, may not extend to private companies, which may not be trusted and have goals and objectives that may not be aligned with public expectations [216]. Thirdly, public sector entities have several competing priorities that may undermine a government's ability to oversee the partnership effectively. A public sector entity may have difficulty in reconciling the objective of successful development of a new product or service, the obligation to protect the rights of individuals and patients and the wider responsibility to regulate all the operations of a private sector partner effectively.

Fourthly, there is often concern that the contributions of the public sector and the community (technology, data, funding, expertise, testing sites) are not considered when allocating ownership rights (if any) to a technology between the public and private sector and in setting the price of such technologies or the rules under which the technology is used [216]. If the public sector and communities make significant contributions to a partnership but are not full beneficiaries, such collaborations may be considered exploitative.

9.3.3 Governance and oversight of large technology companies

Large technology companies, especially those located in China and the USA, are expected to play a central role in the development and deployment of AI for health, through partnerships, in-house development of AI or acquisition of other companies. The role and involvement of these companies raises further considerations for oversight of the private sector. Large technology companies, of which there are only a few, wield significant power in the field of AI because of their human, economic and technical resources, the data accumulated from their products and services, the political influence they may be able to exert through their relationships and partnerships with governments and their staff (see below) and their ability to use their platforms to introduce products and services to large numbers of users, who are regularly connected to their platforms.

Over time, large technology companies may develop even more diversified products and services. Google is developing a range of diagnostic applications that are still being examined for safety and efficacy, and its parent holding company, Alphabet, has launched a new health insurance service that will work in partnership with SwissRe [318].

Companies may also launch products and services that could compete with, replace or introduce a function or process that is usually managed by a government. Tencent has introduced an application that uses information voluntarily supplied by individuals to determine the type of health-care provider a patient should consult, partly to resolve a practice in China whereby patients use their own research or intuition to seek medical advice from specialists in areas unrelated to their condition.[8] The growth of telemedicine is providing opportunities for company-owned platforms to move patients to their platforms, and they are enrolling doctors to provide services via the platform. For example, Tencent WeDoctor, which works with the Government, has enrolled at least 240 000 providers onto its platform and also 2700 hospitals and 15 000 pharmacies. At least 27 million monthly users consult the "health-care collaboration platform" for an AI-guided or a remote consultation. Users are then matched with the appropriate specialist in the health-care system [319]. This could mean that, in the long term, governments might not so much regulate companies that provide such services but might depend on them to fill gaps and manage parts of the health-care system. Technology companies may supply the infrastructure for operation of health-care services, which also creates dependence of

8 Presentation by Alexander Ng, Tencent, 27 August 2020, to the WHO Expert Group on AI for health.

governments on the services and capabilities of the companies, rather than regulating the industry to serve the needs of the government and the public.

As noted above, technology companies have begun to issue guiding principles for the use of AI; however, they are sometimes viewed as "ethics washing", may create a gap in responsibility (assigning responsibility for retrospective harm), do not involve the public in their development and may be administered in a way that is not transparent to the public or to governments, with no involvement of the public or an independent authority for oversight of adherence to the principles.

9.3.4 Provision of health care by the private sector outside the health-care system

The proliferation of AI applications for health outside the health-care system may extend access to some health-care advice; however, such applications raise new questions and concerns. An application may be developed without appropriate reference to clinical standards; it may not be user friendly, especially for follow-up services or procedures; patient safety may be compromised if individuals are not connected to health-care services, such as lack of assistance to individuals with suicidal ideation who use an AI chatbot; the efficacy of applications such as chatbots that may not have been tested properly may be inadequate; and applications may not meet the standards of privacy required for sensitive health data [319]. As such applications are not necessarily labelled as health-care services and may not even be known to governments, the overall quality of health care could be compromised, and people with no other options may be relegated to subpar services. Governments should identify these applications, set common standards and regulations (or even prevent some applications from being deployed to the public) and ensure that individuals who use the applications retain access to appropriate health-care services that cannot be provided online.

9.3.5 An enabling environment for effective governance of the private sector

Appropriate governance of the private sector must overcome a number of hurdles. One is the power of many of the companies involved in delivering AI for health care. Many of them employ former government officials and regulators, who are asked to lobby and influence policy-makers and regulators charged with overseeing the use of AI for health care. This can affect the ability of governments to act independently of companies.

A second challenge is that many of the technologies developed by companies are increasingly difficult to evaluate and oversee, partly because of their growing complexity, including the use of black-box algorithms and deep learning methods. The growing complexity has encouraged both governments and companies to consider models of "co-regulation", whereby each party relies on the other to assess and regulate a technology. While such models of oversight may assist governments in understanding a technology, they may limit the government's exercise of independent judgement and encourage them to trust that companies are willing to strictly self-regulate their practices.

Improving governance of the private sector in other ways will require more independent in-house expertise and information so that governments can evaluate and regulate company practices effectively. Thus, capacity-building of government regulators and transparency will both play roles in improving government oversight of the private sector. Such measures could include greater transparency of the data collected and used by private companies, how ethical and legal principles are integrated into company operations and how products and services perform in practice, including how algorithms change over time.

Recommendations

1. Governments should ensure that the growing provision of health-related services through online platforms that are not associated with the formal health-care system is identified, regulated (including standards of privacy protection guaranteed within health-care systems) and avoided for areas of health care in which the safety and care of patients cannot be guaranteed. Governments should ensure that patients who use such services also have access to appropriate formal health-care services when required.

2. Governments should consider adopting models of co-regulation with the private sector to understand an AI technology, without limiting independent regulatory oversight. Governments should also consider building their internal capacity to effectively regulate companies that deploy AI technologies and improve the transparency of a company's relevant operations.
3. Governments should consider establishing dedicated teams to conduct objective peer reviews of software and system implementation by examining safety and quality or general system functionality (fitness for purpose) without requiring review or approval of a code.
4. Governments should consider which aspects of health-care delivery, financing, services and access could be supplied by companies, how to hold them accountable and which aspects should remain the obligation of governments.
5. Public-Private Partnerships (PPPs) that develop or deploy AI technologies for health should be transparent (including in the terms and conditions of any agreement between a government and a company) through meaningful engagement by the public. Such partnerships should prioritize protection of individual and community rights and governments should seek ownership rights to products and services so that the outcomes of the PPP are affordable and available to all.
6. Companies must adhere to national and international laws and regulations on the development, commercialization and use of AI for health systems, including legally enforceable human rights and ethical obligations, data protection laws, measures to ensure appropriate informed consent and privacy.
7. Companies should invest in measures to improve the design, oversight, reliability and self-regulation of their products. Companies should also consider licensing or certification requirements for developers of "high-risk" AI, including AI for health.
8. Companies should ensure the greatest possible transparency in their internal policies and practices that implicate their legal, ethical and human rights obligations as established under the UN Guiding Principles on Business and Human Rights. They should be transparent about how those ethical principles are implemented in practice, including the outcomes of any actions taken to address violations of such principles.

9.4 Governance of the public sector

Use of AI in the public sector has increased recently, although it lags behind adoption by the private sector. In 2019, OECD identified 50 countries that have launched or are planning to launch national AI strategies, of which 36 plan to or have issued separate strategies for public sector AI [320]. In 2017, the United Arab Emirates was the first country in the world to have a designated minister for AI, which has resulted in increased use of AI in the health-care system, such as "pods" to detect early signs of illness, AI-enabled telemedicine and use of AI to detect diabetic retinopathy [321]. Although use of AI has increased in the public sector, a review of nearly 1700 studies found only 59 on use of AI in the public sector [320]. There is no comprehensive account of how governments are advancing the use of AI or integrating it into health care. The OECD identified six broad roles for governments in AI, as a:

– financier or direct investor in AI technologies in both the public and the private sector;
– "smart buyer" and co-developer, including PPPs and other forms of collaboration with companies;
– regulator or rule-maker;
– convenor and standard setter;
– data steward; and
– user and services provider.

This section briefly addresses how governments should use AI ethically as investors in AI technologies, as smart buyers and/or co-developers and as users and service providers. It also addresses concern about ethics and human rights with increased use of AI to manage social protection and welfare, programmes that often directly influence access to health-care services and indirectly affect human health and well-being.

9.4.1 Assessing whether AI is necessary and appropriate for use in the public sector

As for any use of AI by health professionals, governments must assess whether an AI technology is necessary and appropriate for the intended use and can be used according to its laws. The assessment could include an evaluation of whether use of AI is appropriate. In India, the Government's internal think tank, Niti Aayog, has proposed constitution of an ethics committee to review procurement of AI in the public sector. According to a draft proposal released in 2020, the committee "may be constituted for the procurement, development, operations phase of AI systems and be made accountable for adherence to the Responsible AI principles" [322]. A requirement that both ministries of health and public and private health-care providers observe legal and ethical standards in the procurement of AI can encourage appropriate design of AI technologies and provide a safeguard against harm.

The Government of the United Kingdom has established an analytical framework for use of AI [323], which consists of the following: whether the available data contain the required information; if it is ethical and safe to use the data and consistent with the Government's data ethics framework; if there are sufficient data for training AI; whether the task is too large or repetitive for a human to undertake without difficulty; and whether AI will provide information that a team could use to achieve real-world outcomes.

9.4.2 Accountability through transparency and participation

Governments are increasingly required to disclose the use of algorithms in services and operations in order to promote accountability for the use of AI, and many data protection laws require that decisions not be taken solely by automated systems and that use of automated decision-making be prevented in certain contexts. In France, the Government is required to provide a general explanation of how any algorithm it uses functions, personalized explanations of decisions issued by algorithms, justification for decisions and publication of the source code and other documentation about the algorithms [320].

In general, there is growing expectation that governments will be transparent about their use of AI, including whether they are investing in AI, engaged in partnerships with companies or developing AI independently in state-owned enterprises or government agencies. It is also expected that governments will be transparent about any harm caused by use of AI and the measures taken to redress any harm. A review conducted by the United Kingdom Committee on Standards in Public Life found that the British Government (during the period examined) had not met established principles of openness and noted that "under the principle of openness, a current lack of information about government use of AI risks undermining transparency" [324].

Yet, transparency may not be sufficient to ensure that government use of algorithms will not result in undue harm, especially for marginalized communities and populations. Greater public participation by a wide range of stakeholders is necessary to ensure that decisions about the introduction of an AI system in health care and elsewhere are not taken only by civil servants and companies but are based on public participation of a wider range of stakeholders, including representatives of public interest groups and leaders of vulnerable groups that are often not involved in making such decisions. Their perspectives should be obtained before and not only after identification of an adverse effect, which is too late.

9.4.3 Appropriate collection, stewardship and use of data

The collection, storage and use of data according to ethical and legal standards also applies to governments. Government use of data is prone to abuse, whether through the sale or provision of data

to private companies that violates the public trust or sharing data obtained or collected for health-care purposes in other government programmes, including enforcement of immigration laws or criminal justice. Such health data, which often include information on location or behaviour, can then be used to infringe on civil liberties directly. These uses of data undermine trust in the health-care system and the willingness of individuals to provide data and use AI technologies that are intended to improve the administration of health care and medicine.

Governments also face risks of bias in data that are collected for the development of AI for use in the public sector. The obligation of the public sector to remain objective may be undermined, as the "prevalence of data bias risks embedding and amplifying discrimination in everyday public sector practice" [325]. The review of use of AI in the public sector in the United Kingdom also found that "data bias is an issue of serious concern, and further work is needed on measuring and mitigating the impact of bias" [324].

9.4.4 Risks and opportunities in use of AI for provision of public services and social protection

Governments have used AI to provide public services, including assessment of whether an individual qualifies for certain services, in what is known generally as the "digital welfare state". Thus, digital data and technologies are used to automate, predict, identify or disqualify potential recipients of social welfare. While some have championed this use of AI as a means of eliminating redundant and repetitive tasks that both saves resources and gives government employees more time to address more difficult issues [325], there is concern that the digital welfare state could undermine access to social services and welfare and especially affect poor and marginalized populations. According to a report by the United Nations Special Rapporteur on extreme poverty and human rights, the digital welfare state could become a "digital dystopia", constricting budgets intended for the provision of services, limiting those who qualify for government services, creating new conditionality and introducing new sanctions to discourage the use of services [326]. The report also notes that administering a welfare state through a digital ecosystem can exacerbate inequality, as many poor and marginalized individuals do not have adequate access to online services [326]. Although the report does not discuss use of AI to provide or refuse health-care services, such use could affect the provision of health care in the public sector or, for example, the provision of health insurance through the public or private sector.

Recommendations

1. Governments should conduct transparent, inclusive impact assessments before selecting or using any AI technology for the health sector and regularly during deployment and use. This should consist of ethics, human rights, safety, and data protection impact assessments. Governments should also define legal and ethical standards for procurement of AI technologies and require public and private health-care providers to integrate those standards into their procurement practices.
2. Governments should be transparent about the use of AI for health, including investment in use, partnerships with companies and development of AI in state-owned enterprises or government agencies, and should also be transparent about any harm caused by use of AI.
3. Governments and national health authorities should ensure that decisions about introducing an AI system for health care and other purposes are taken not only by civil servants and companies but with the democratic participation of a wide range of stakeholders and in response to needs identified by the public health sector and patients. They should include representatives of public interest groups and leaders of marginalized groups, who are often not considered in making such decisions.

4. Governments should develop and implement ethical, legally compliant principles for the collection, storage and use of data in the health sector that are consistent with internationally recognized data protection principles. In particular, governments should take steps to avoid risks of bias in data that are collected and used for development and deployment of AI in the public sector.
5. Governments should ensure that any use of AI to facilitate access to health care is inclusive, such that uses of AI do not exacerbate existing health and social inequities or create new ones.

9.5 Regulatory considerations

The largest national regulatory agencies, such as the Food and Drug Administration in the USA, have been developing guidance and protocols to ensure the safety and efficacy of new AI technologies; however, other regulatory agencies may have neither the capacity nor the expertise to approve use of such devices. A WHO working group has been formed to address regulatory considerations for the use of AI for health care and drug development and will issue a report and recommendations in 2021. The present guidance identifies several ethical concerns that could be addressed by regulatory agencies and the challenges that could arise.

9.5.1 Does regulation stifle innovation?

It is commonly asserted that stringent regulations will limit innovation and deprive health-care systems, providers and patients of beneficial innovations. A balance must be struck between protecting the public and promoting growth and innovation [159]. Use of AI for health is still new and often untested, and policy-makers and regulators must consider numerous ethical, legal and human rights issues. For example, regulators must identify those applications and AI-based devices that may be best described as "snake oil", a euphemism for deceptive marketing, health-care fraud or a scam, which either misrepresents what an application can do, provides misinformation or persuades vulnerable individuals to follow health advice that may be contrary to their well-being [327].

Applications that provide no therapeutic or health benefit might be introduced solely for collecting health and biological data for use in commercial marketing or to encourage patients to pay for irrelevant or unproven health interventions [328]. For example, an academic obtained data from 300 000 Facebook users who were told that the data were for a "psychological test". Their data and data from an estimated 50 million other users linked to them (Facebook "friends") were then sold to Cambridge Analytica, which used them to build a software program to predict and influence choices at the ballot box [329]. Such malicious use of data collected nominally for academic or health purposes could expose health systems, health providers and companies that provide health-related AI services to significant risk.

Regulation could differ according to risk, such that those who are especially vulnerable, including people with mental illness, children and the elderly, are protected from misinformation and bad advice from health applications that exploit rather than assist such individuals [159]. People living in resource-poor settings, in countries with inadequate resources to regulate and monitor adverse consequences of AI applications and with diseases that result in marginalization and discrimination, such as HIV/AIDS or tuberculosis, also require greater protection and oversight by regulatory agencies than users of applications for lifestyle or wellness.

9.5.2 Transparency and explainability of AI-based devices

The black box of machine learning creates challenges for regulators, who may be unable to fully assess new AI technologies because the standard measures used to assess the safety and efficacy of medical technologies and scientific understanding and clinical trials are not appropriate for black-box medicine [255]. Complex algorithms are difficult for regulators to understand (partly because of lack of expertise in regulatory agencies) and difficult for developers to explain.

Improving the scientific understanding (explainability) of an algorithm is considered necessary to ensure that regulators (and clinicians and patients) understand how a system arrives at a decision. Explainability is also a requirement of the EU's GDPR and is being introduced into legislation in other countries experiencing proliferation of AI for health care and other fields [116]. It has been argued that, if a trade-off is to be made between transparency and accuracy, transparency should predominate. This requirement may, however, not be possible or even desirable in the medical context. While it is often possible to explain why a specific treatment is the best option for a specific condition, it is not always possible to explain how that treatment works or its mechanism of action, because medical interventions are sometimes used before their mode of action is understood.

Trust in decisions and expert recommendations depends on the ability of experts to explain why a certain system is the best option for achieving a clinical goal. Such explanations should be based on reliable evidence of the superior accuracy and precision of an AI system over alternatives. The evidence should be generated by prospective testing of the system in randomized trials and not their performance against existing datasets in a laboratory.

Understanding how a system arrives at judgements may be valuable for a variety of reasons, but it should not take precedence over or replace sound, prospective evidence of the system's performance in prospective clinical trials. Explanations of how a system arrives at a particular decision could encourage use of machine-learning systems for purposes for which they are not well suited, as the models created by such systems are based on associations among a wide range of variables, which are not necessarily causal. If the associations are causal, practitioners might rely on them to make decisions for which the system has not been tested or validated. Requiring every clinical AI decision to be "explainable" could also limit the capacity of AI developers to use AI technologies that outperform older systems but which are not explainable [116].

Clinical trials provide assurance that unanticipated hazards and consequences of AI-based applications can be identified, addressed and avoided entirely, and additional testing and monitoring of an approved AI device can demonstrate its performance and any changes that may occur after it has been approved. Clinical trials, especially those carried out with diverse populations, can also indicate whether an AI technology is biased against certain sub-groups, races or ethnicities (see below). Clinical trials may not, however, be appropriate because of their cost, because it takes a long time to conduct a trial properly, because the validity of the results may be called into question if an algorithm is expected to change over time with new data, and because AI-based technologies and products are increasingly personalized to smaller populations and therefore more difficult to test with enough individuals [255].

Clinical trial designs and statistical analysis strategies should be re-evaluated, and innovation should be encouraged in these areas of AI validation. While AI should properly be validated in clinical trials or other applicable ways, AI itself could potentially allow even more accurate trials of device or drug effectiveness with smaller patient populations through enhanced patient-trial matching, data analytics efficiency and other approaches. This might become relevant during the COVID-19 pandemic as recruitment and access to health-care facilities is challenged.

Regulators could introduce "lighter premarket scrutiny" in the place of clinical trials for AI technologies for health, by assessing the safeguards put in place by developers, the quality of the data used, development techniques, validation procedures and "robust post-market oversight". This might, however, be difficult to implement in practice, especially post-market oversight of novel algorithms [255], and may be too late to prevent harm to people who are especially vulnerable, such as those who have no access to a health-care provider who could protect them from a misguided diagnosis or advice. The transparency of the initial dataset could be improved, including the provenance of the data and how they were processed, as could the transparency of the system architecture [115]. Such transparency would allow others to validate an AI technology independently and increase the trust of users.

While greater transparency of the components of an AI system, including its source code, data inputs and analytical approach, can facilitate regulatory oversight, some transparency may misplace focus. Reviewing lines of code would be time-consuming and unlikely to be informative in comparison with the performance, functionality and accuracy of the system both before and after it is integrated into a health-care system.

9.5.3 Addressing bias

Regulatory agencies should create incentives to encourage developers to identify and avoid biases. One example is the addition of measures to a precertification programme hosted by the US Food and Drug Administration, the Digital Health Innovation Action Plan [330]. The programme already assesses medical software on the basis of criteria of excellence, including quality. The criteria for quality and other criteria set by regulatory agencies could include the risk of bias in training data [330]. Robust post-marketing surveillance to identify biases in machine-learning algorithms, including in collaboration with providers and communities likely to be affected by biased algorithms, could improve regulatory oversight.

9.5.4 Ethical considerations for LMIC and HIC with poor health outcomes

LMIC often have insufficient regulatory capacity, so that they are unable to assess the safety and efficacy of new technologies. Regulatory agencies in LMIC could consider either relying on regulatory approval of AI technologies in HIC or use of collaborative registration procedures to ensure that new technologies are appropriate for use. Global harmonization of regulatory standards would ensure that all countries benefit from rigorous testing, transparent communication of outcomes and monitoring of a technology's performance. International harmonization of regulatory standards, based on those of HIC, or reliance on other regulatory agencies or the assurances of product developers is founded on the assumption that the criteria used to develop or assess a new technology in HIC is appropriate for LMIC contexts and populations. This may not be the case, and it is likely that AI health technologies cannot be transposed between divergent settings, including between LMIC and HIC [115]. This may be due not only to the types of data used to train the algorithm but also to the assumptions and definitions used in developing an AI technology, such as what constitutes "healthy", which may be defined by a small group of developers located in one company or country and validated by regulators in HIC with no consideration of whether the assumptions are appropriate for LMIC [183].

Regulators may also make assumptions about the context in which an AI technology was introduced. AI technologies may have "contextual bias", whereby the algorithms may not recommend safe, appropriate or cost-effective treatments for low-income or low-resource settings [193] or for countries that have resources but in which segments of the population still have poor health outcomes, as is often the case in some HIC. The developer of a technology for a high-income setting in which most of the population have good health outcomes may neither anticipate nor build an AI technology to anticipate differences from LMIC settings or from other HIC with poor health outcomes, and a regulator, even if it requires prospective clinical trials, may not require data on how the technology operates in LMIC or certain high-income settings.

While the transparency of the data used to train algorithms, the context in which an algorithm is trained and other material assumptions are necessary, they may only delay use of an AI technology, thus avoiding harm, but not bestow any benefit. Improving the performance and use of AI technologies in LMIC and certain HIC and ensuring that the technologies are adapted to reality will require different incentives, approaches and developers of technologies that are appropriate for all people [193].

Recommendations

1. Governments should introduce and enforce regulatory standards for new AI technologies to promote responsible innovation and to avoid the use of harmful, insecure or dangerous AI technologies for health.

2. Government regulators should require the transparency of certain aspects of an AI technology, while accounting for proprietary rights, to improve oversight and assurance of safety and efficacy. This may include an AI technology's source code, data inputs and analytical approach.
3. Government regulators should require that an AI system's performance be tested and sound evidence obtained from prospective testing in randomized trials and not merely from comparison of the system with existing datasets in a laboratory.
4. Government regulators should provide incentives to developers to identify, monitor and address relevant safety- and human rights-related concerns during product design and development and should integrate relevant guidelines into precertification programmes. Regulators should also mandate or conduct robust marketing surveillance to identify biases.

9.6 Policy observatory and model legislation

As AI plays a more prominent role in health systems, governments are introducing national policies and laws to govern its use in health. To ensure that such laws and policies address the ethical concerns and the opportunities associated with use of AI, the OECD launched a policy observatory in 2020 that "aims to help countries enable, nurture and monitor the responsible development of trustworthy artificial intelligence systems for the benefit of society" [331].

WHO supports such initiatives and, on the basis of the ethical principles and findings outlined in this document, is exploring collaboration with the OECD on a policy observatory to identify and analyse relevant policies and laws. It is critical that WHO collaborate with other well-placed intergovernmental organizations with wider membership, including of LMIC, such as other United Nations agencies. WHO may also consider issuing model legislation as a reference for governments to develop their own laws to ensure appropriate protection, regulations, rules and safeguards to build the trust of the general public, providers and patients in the use of AI in health-care systems, and, for example, for the management of data and information in ways that improve the accuracy and utility of AI while not compromising privacy, confidentiality or informed consent.

Recommendations

1. WHO should work in a coordinated manner with appropriate intergovernmental organizations to identify and formulate laws, policies and best practices for ethical development, deployment and use of AI technologies for health.
2. WHO should consider issuing model legislation to be used as a reference for governments that wish to build an appropriate legal framework for the use of AI for health.

9.7 Global governance of artificial intelligence

AI is playing an ever-expanding role worldwide. AI has already contributed US$ 2 trillion to global gross domestic product, which could rise to more than US$ 15 trillion by 2030 [332]. The importance of AI can also be measured by the positive or negative role it might play in achievement of the Sustainable Development Goals. According to one study, AI could enable accomplishment of 134 of the targets but inhibit achievement of 59 targets [6].

Ethical principles, regulatory frameworks and national laws on AI continue to proliferate, providing a form of governance; however, the ethical principles and guidance on adherence to international human rights obligations related to AI remain nascent and differ widely among countries, in the public and the private sector and between governments and companies; the platforms of several companies boast more users or subscribers than those of the most populous countries. Thus, company standards influence the control of many AI technologies, including those used in health care.

With the increase in AI standards and laws around the world and diffusion of how and where AI ethics is managed, additional international oversight and enforcement may be necessary to ensure convergence on a core set of principles and requirements that meet ethical principles and human rights

obligations. Otherwise, the short-term economic gains that could be made with AI could encourage some governments and companies to ignore ethical requirements and human rights obligations and engage in a "race to the bottom".

First, technical advice from and the engagement of WHO and other intergovernmental organizations such as the Council of Europe, OECD and UNESCO and respect for ethical principles and human rights standards can ensure that companies and governments both move towards common high standards [333]. In the domain of global health, this will also require that major global health bodies, such as WHO, the Global Fund to Fight AIDS, Tuberculosis and Malaria, United Nations development agencies and foundations, agree on a common position about the risks associated with these technologies and clearly commit themselves to adherence to human rights and ethical standards as a core principle of all strategies and guidance [333].

Secondly, global governance could strengthen the voice and role of LMIC, which are less involved in developing AI technologies or in setting international principles. LMIC also lag in use of AI, including in health, partly because of the enduring digital divide, and may not yet have the capacity to regulate use of AI. Thus, global governance could improve access to information and communication and digital technologies in LMIC, guide LMIC governments in accurate assessment of the benefits and risks of AI technologies and hold companies accountable for their practices in LMIC.

Thirdly, global governance could ensure that all governments can adapt to the changes that will be wrought as these technologies become ever more sophisticated and powerful. Independent scientific advice and evidence will be necessary as AI technologies evolve and are translated into policy guidance. For the use of AI for health, it is critical that global health agencies promote only those AI technologies that have been rigorously tested and validated as health interventions by an appropriate authority, such as WHO, and assessed for risks [333].

Global governance of use of AI for health will consist partly of adapting governance structures, including the policies and practices of global health agencies, treatment guidelines issued by WHO and global agreements to meet certain health objectives, such as eliminating HIV and AIDS by 2030. Furthermore, global standards should be set for all ethical concerns of AI for health, such as impacts on labour, data governance, privacy, ownership and autonomous decision-making.

As for the use of many other health technologies, nongovernmental organizations and community groups will play critical roles in ensuring that human rights obligations and ethical principles are considered from the onset of decision-making and respected in practice and that governments and companies introduce appropriate safeguards to prevent and respond to any risks and swiftly redress any negative consequences of the use of AI. Civil society and affected communities should participate in the design of AI technologies, and international organizations should work with nongovernmental organizations and affected populations to develop and mainstream guidance for governments and companies.

Several efforts have been made to improve global governance of AI, including the joint initiative of the governments of Canada and France to establish the Global Partnership on AI in June 2020, which now comprises 19 countries. It is intended to convene global AI experts and provide guidance on AI topics, including the future of work, data and privacy [334]. Its first summit was held in December 2020 [335].

Such welcome bilateral and multilateral initiatives should feed into global processes based on the perspectives of all countries. For example, the United Nations Secretary-General's Roadmap for digital cooperation [336] recommended in 2019

> creating a strategic and empowered multi-stakeholder high-level body, building on the experience of the existing multi-stakeholder advisory group, which would address urgent issues, coordinate follow-up action on Forum discussions and relay proposed policy approaches

and recommendations from the Forum to the appropriate normative and decision-making forums.

Such a multi-stakeholder body would contribute to the wider governance and standard-setting required for AI and provide means for addressing many of the challenges and questions related to the ethics and governance of the use of AI for health.

Recommendations

1. Governments should support global governance of AI for health to ensure that the development and diffusion of AI technologies is in accordance with the full spectrum of ethical norms, human rights protection and legal obligations.
2. Global health bodies such as WHO, Gavi, the Vaccines Alliance, the Global Fund to Fight AIDS, Tuberculosis and Malaria, Unitaid and major foundations should commit themselves to ensuring that adherence to human rights obligations, legal safeguards and ethical standards is a core obligation of all strategies and guidance.
3. International agencies, such as the Council of Europe, OECD, UNESCO and WHO, should develop a common plan to address the ethical challenges and the opportunities of using AI for health, for example through the United Nations Interagency Committee on Bioethics. The plan should include providing coherent legal and technical support to governments to comply with international ethical guidelines, human rights obligations and the guiding principles established in this document.
4. Governments and international agencies should engage nongovernmental and community organizations, particularly for marginalized groups, to provide diverse insights.
5. Civil society should participate in the design and use of AI technologies for health as early as possible in their conceptualization.

References

1. Report of the Secretary-General on SDG progress. Special edition. New York City (NY): United Nations; 2019 (https://sustainabledevelopment.un.org/content/documents/24978Report_of_the_SG_on_SDG_Progress_2019.pdf, accessed 8 November 2020).
2. Timmermans S, Kaufman R. Technologies and health inequities. Ann Rev Sociol. 2020;46:583-602.
3. Report of the Special Rapporteur on the Promotion and protection of the right to freedom and expression. United Nations General Assembly. 73rd Session (A/73/348). New York City (NY): United Nations; 2018 (https://undocs.org/pdf?symbol=en/A/73/348; accessed 7 January 2021).
4. Recommendation of the Council on Artificial Intelligence (OECD Legal Instruments. OECD/LEGAL/O449). Paris: Organization for Economic Co-operation and Development; 2019 (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#mainText, accessed 2 December 2020).
5. Hao K. What is machine learning? Machine-learning algorithms find and apply patterns in data. And they pretty much run the world. MIT Technology Revies,17 November 2017 (https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/, accessed 28 August 2020).
6. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun. 2020;11:233.
7. Flynn L. When AI is watching patient care: Ethics to consider. Bill of Health, 18 February 2020 (https://blog.petrieflom.law.harvard.edu/2020/02/18/when-ai-is-watching-patient-care-ethics-to-consider/, accessed 13 August 2020).
8. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? BMJ Glob Health. 2018;3:e000798.
9. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet. 2020;395:1579-86.
10. Miller RA, Schaffner KF, Meisel A. Ethical and legal issues related to the use of computer programs in clinical medicine. Ann Intern Med. 1985;102:529-36.
11. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin. 2019;69[2]:127-57.
12. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li T. Automatic detection of Mycobacterium tuberculosis using artificial intelligence. J Thorac Dis. 2018;10[3]:1936-40.
13. Mandavilli A. These algorithms could bring an end to the world's deadliest killer. New York Times. 20 November 2020 (https://nyti.ms/2KnQPu5, accessed 19 January 2021).
14. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digital Health. 2019;1:6.
15. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15[11]:1002686.

16. Bejnordi BE, Veta M, van Diest PJ, van Ginneken P, Karssemeijer N, Litjens J et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318[22]:2199-210.

17. Alsharqi M, Woodward WJ, Mumith JA, Markham DC, Upton R, Leeson P. Artificial intelligence and echocardiography. Echo Res Pract. 2018;5[4]:R115-25.

18. Collis F. Using artificial intelligence to detect cervical cancer. NIH Director's Blog, 17 January 2019 (https://directorsblog.nih.gov/2019/01/17/using-artificial-intelligence-to-detect-cervical-cancer/, accessed 15 February 2021).

19. Innovative, affordable screening and treatment to prevent cervical cancer. Geneva: Unitaid; 2021 (https://unitaid.org/project/innovative-affordable-screening-and-treatment-to-prevent-cervical-cancer/#en, accessed February 2021).

20. Fan R, Zhang N, Yang L, Ke J, Zhao D, Cui Q. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. Sci Rep. 2020;10:14457.

21. Yan Y, Zhang JW, Zang GY, Pu J. The primary use of artificial intelligence in cardiovascular diseases: What kind of potential role does artificial intelligence play in future medicine? J Geriatr Cardiol. 2019;16[8]:585-91.

22. Chaki J, Thillai Ganesh S, Cidham SK, Theertan SA. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review. J King Saud Univ Comput Inf Sci. 2020 (https://doi.org/10.1016/j.jksuci.2020.06.013, accessed 12 September 2022).

23. Singh J. Artificial intelligence and global health: opportunities and challenges. Emerg Topics Life Sci. 2019;3:10.

24. The Topol review: Preparing the healthcare workforce to deliver the digital future. London: National Health Service; 2019 (https://topol.hee.nhs.uk/, accessed 23 August 2020).

25. Hollander JE, Carr BG. Virtually perfect? Telemedicine for COVID-19. N Engl J Med. 2020;382:1679-81.

26. Mou M. COVID-19 gives boost to China's telemedicine industry. Wall Street Journal, 22 October 2020 (https://www.wsj.com/articles/covid-19-gives-boost-to-chinas-telemedicine-industry-11603379296, accessed 3 February 2021).

27. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digit Health. 2019;5:2055207619871808.

28. Dennis AR, Kim A, Rahimi M, Ayabakan S. User reactions to COVID-19 screening chatbots from reputable providers. J Am Med Informatics Assoc. 2020;27[11]:1727-31.

29. Roski J, Chapman W, Heffner J, Trivedi R, Del Fiol G, Kukafka R et al. How artificial intelligence is changing health and health care. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019 (https://nam.edu/artificial-intelligence-special-publication/, accessed 19 July 2020).

30. Marr B. The incredible ways in which artificial intelligence is now used in mental health. Forbes, 3 May 2019 (https://www.forbes.com/sites/bernardmarr/2019/05/03/the-incredible-ways-artificial-intelligence-is-now-used-in-mental-health/?sh=7806594ad02e, accessed 17 May 2020).

31. Gamble A. Artificial intelligence and mobile apps for mental healthcare: a social informatics perspective. Aslib J Inf Manag. 2020;72[4]:509-23.

32. What is "biosurveillance"? The COVID-19 measures getting under our skin. Amsterdam: Digital Freedom Fund, 28 May 2020 (https://medium.com/digital-freedom-fund/what-is-biosurveillance-c8bffe70d16f, accessed 17 October 2020).

33. Vincent JL, Moreno R, Takala J, Willatts S, De Mendança A, Bruining H et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996;22:707-10.

34. Khanam V, Tusha J, Abkouh DT, Al-Janabi L, Tegeltija V, Kumar S. Sequential organ failure assessment score in patients infected with SARS COV-2. Chest. 2020;158[4]:A602.

35. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A continuous acuity score for critically ill patients using clinically interpretable deeplearning. Sci Rep. 2019;9:1879.

36. Shea GP, Solomon CA. Triage in a pandemic: Can AI help ration care? Knowledge@Wharton, 27 March 2020. Philadelphia (PA): University of Pennsylvania (https://knowledge.wharton.upenn.edu/article/triage-in-a-pandemic-can-ai-help-ration-access-to-care/, accessed 3 December 2020).

37. Babic B, Cohen IG, Evgeniou T, Gerke S, Trichakis N. Can AI fairly decide who gets an organ transplant. Harvard Business Review, 1 December 2020. Cambridge (MA): Harvard Business Publishing (https://hbr.org/2020/12/can-ai-fairly-decide-who-gets-an-organ-transplant, accessed 3 December 2020).

38. Raza S. Artificial intelligence for genomic medicine. Cambridge: PHG Foundation, University of Cambridge; 2020 (https://www.phgfoundation.org/documents/artifical-intelligence-for-genomic-medicine.pdf, accessed 11 December 2020).

39. Fleming N. How artificial intelligence is changing drug discovery. Nature Spotlight: Biopharmaceuticals, 30 May 2018 (https://www.nature.com/articles/d41586-018-05267-x, accessed 13 October 2020).

40. New Ebola treatment using artificial intelligence. San Francisco (CA): Atomwise; 2015 (https://www.atomwise.com/2015/03/24/new-ebola-treatment-using-artificial-intelligence/, accessed February 2020).

41. Metz C. London AI lab claims breakthrough that could accelerate drug discovery. The New York Times, 30 November 2020 (https://nyti.ms/2VfKkvA, accessed 14 December 2020).

42. Low LA, Mummery C, Berridge BR, Austin CP, Tagle DA. Organs-on-chips: into the next decade. Nat Rev Drug Discov. 2020 (https://doi.org/10.1038/s41573-020-0079-3, accessed April 2021).

43. Artificial intelligence: How to get it right. London: National Health Service; 2019 (https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf, accessed 2 August 2020).

44. WHO guidelines on ethical issues in public health surveillance. Geneva: World Health Organization; 2017 (https://apps.who.int/iris/bitstream/handle/10665/255721/9789241512657-eng.pdf, accessed 13 September 2022).

45. Micro-targeting. London: Privacy International; 2021 (https://privacyinternational.org/learn/micro-targeting, accessed 13 January 2021).

46. Smart cities. London: Privacy International; 2021 (https://privacyinternational.org/learn/smart-cities, accessed 15 January 2021).

47. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012-4.

48. Privacy International and the International Committee for the Red Cross. The humanitarian metadata problem: Doing no harm in the digital era. London: Privacy International and ICRC; 2018 (https://privacyinternational.org/sites/default/files/2018-12/The%20Humanitarian%20Metadata%20Problem%20-%20Doing%20No%20Harm%20in%20the%20Digital%20Era.pdf, accessed 6 February 2021).

49. Cho A. Artificial intelligence systems aim to sniff out signs of COVID-19 outbreaks. Science, 12 May 2020 (https://www.sciencemag.org/news/2020/05/artificial-intelligence-systems-aim-sniff-out-signs-covid-19-outbreaks#, accessed August 2020)

50. Hswen Y, Brownstein JS. Real-time digital surveillance of vaping-induced pulmonary disease. NEJM. 2019;381:1778-80.

51. White RW, Wang S, Pant A, Harpaz R, Shukla P, Sun W et al. Early identification of adverse drug reactions from search log data. J Biomed Informatics. 2016;59:42-8.

52. Precision FDA: Gaining new insights by detecting adverse event anomalies using FDA Open Data. Silver Spring (MD): US Food and Drug Administration; 2020 (https://precision.fda.gov/challenges/9, accessed September 2020).

53. Whitelaw S, Mamas MA, Topol E, Van Spall GC. Applications of digital technology in COVID-19 planning and response. Lancet Digital Health. 2020;2 e435-40.

54. Bullock J, Luccioni A, Pham KH, Nga Lam CS, Luengo-Oroz M. Mapping the landscape of artificial intelligence applications against COVID-19. J Artificial Intell Res. 2020;69:807-45.

55. Toh A. Big Data could undermine the COVID-19 response. Wired, 12 April 2020 (https://www.wired.com/story/big-data-could-undermine-the-covid-19-response/, accessed February 2021).

56. McDonald SM. Ebola: A big data disaster. Privacy, property, and the law of disaster experimentation (CIS Papers 2016.01). Delhi: Centre for Internet and Society; 2016 (https://cis-india.org/papers/ebola-a-big-data-disaster, accessed 20 February 2021).

57. Hao K., Doctors are using AI to triage COVID-19 patients. The tools may be here to stay. MIT Technology Review. 23 April 2020 (https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/, accessed 4 October 2020).

58. AI and control of COVID-19 coronavirus. Strasbourg: Council of Europe; 2020 (https://www.coe.int/en/web/artificial-intelligence/ai-and-control-of-covid-19-coronavirus, accessed 17 September 2020).

59. Ethical considerations to guide the use of proximity tracking technologies for COVID-19 contact tracing. Interim guidance. Geneva: World Health Organization' 2020 (https://www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1, accessed 1 February 2021).

60. Olson, P., Coronavirus reveals limits of AI health tools. Wall Street Journal, 29 February 2020 (https://www.wsj.com/articles/coronavirus-reveals-limits-of-ai-health-tools-11582981201, accessed 5 October 2020).

61. Horowitz BT. Are medical chatbots able to detect coronavirus? Health Tech Magazine, 10 September 2020 (https://healthtechmagazine.net/article/2020/09/are-medical-chatbots-able-detect-coronavirus, accessed 28 October 2020).

62. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1:389-99.

63. Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights: Report of the Secretary General. Geneva: Office of the High Commissioner for Human Rights; 2020 (https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session43/Documents/A_HRC_43_29.pdf, accessed 9 January 2021)

64. Secretary-General Guterres calls for a global reset to recover better, guided by human rights. Geneva: United Nations Human Rights Council; 2021 (https://www.ohchr.org/EN/HRBodies/HRC/Pages/NewsDetail.aspx?NewsID=26769&LangID=E, accessed 3 March 2021).

65. The Toronto Declaration. Protecting the right to equality and non-discrimination in machine learning systems. Amnesty International and Access Now; 2018 (https://www.torontodeclaration.org/declaration-text/english/, accessed 4 June 2020).

66. Addressing the impact of algorithms on human rights. Strasbourg: Council of Europe' 2019 (https://rm.coe.int/draft-recommendation-of-the-committee-of-ministers-to-states-on-the-hu/168095eecf, accessed 16 December 2020).

67. European Convention on Human Rights. Strasbourg: Council of Europe; 2010 (https://www.echr.coe.int/documents/convention_eng.pdf, accessed 6 March 2021)

68. Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine. Strasbourg: Council of Europe; 1997 (https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168007cf98, accessed 3 March 2020).

69. Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data. Strasbourg: Council of Europe; 1981 (https://rm.coe.int/1680078b37, accessed 13 April 2020).

70. Guidelines on artificial intelligence and data protection. Strasbourg: Council of Europe; 2019 (https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8, accessed 13 April 2020).

71. European ethical charter on the use of artificial intelligence in judicial systems and their environment. Strasbourg: Council of Europe; 2018 (https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c, accessed 20 April 2020).

72. General recommendations for the processing of personal data in artificial intelligence. Brussels: Red IberoAmerica de Proteccion de Datos, European Union; 2019 (https://www.redipd.org/sites/default/files/2020-02/guide-general-recommendations-processing-personal-data-ai.pdf, accessed 27 October 2020).

73. Specific guidelines for compliance with the principles and rights that govern the protection of personal data in artificial intelligence projects. Brussels: Red IberoAmerica de Proteccion de Datos, European Union; 2019 (https://www.redipd.org/sites/default/files/2020-02/guide-specific-guidelines-ai-projects.pdf, accessed 27 October 2020).

74. Recommendation CM/Rec [2019]2 of the Committee of Ministers to Member States on the protection of health-related data. Strasbourg: Council of Europe; 2019 (https://www.apda.ad/sites/default/files/2019-03/CM_Rec%282019%292E_EN.pdf, accessed 14 April 2020).

75. African Union Convention on Cyber Security and Personal Data Protection. Addis Ababa: African Union; 2014 (https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection, accessed 19 February 2021).

76. Internet Society, Commission of the African Union. Personal data protection guidelines for Africa. Reston (VA): Internet Society; 2018 (https://www.internetsociety.org/resources/doc/2018/personal-data-protection-guidelines-for-africa/, accessed 19 February 2021).

77. The digital transformation strategy for Africa (2020-2030). Addis Ababa: African Union; 2020 (https://au.int/sites/default/files/documents/38507-doc-dts-english.pdf, accessed 12 February 2021).

78. Recommendations for data and biospecimen governance in Africa. Nairobi: African Academy of Sciences; 2021 (https://www.aasciences.africa/sites/default/files/Publications/Recommendations%20for%20Data%20and%20Biospecimen%20Governance%20in%20Africa.pdf, accessed 26 February 2021).

79. Zeng Y, Lu E, Huangfu C. Linking artificial intelligence principles. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, Honolulu, Hawaii, 2019. Aachen: CEUR Workshop Proceedings; 2019 (https://arxiv.org/ftp/arxiv/papers/1812/1812.04814.pdf, accessed 12 February 2020).

80. OECD legal instruments. Recommendations of the Council on artificial Intelligence. Paris: Organization for Economic Co-operation and Development; 2019 (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL0449, accessed 12 February 2020).

81. Going digital. Making the transformation work for growth and well-being. Paris: Organization for Economic Co-operation and Development; 2019 (https://www.oecd.org/going-digital/ai/principles/, accessed 5 February 2020).

82. Economy, employment, and education in the digital age. What can the G20 do to implement AI principles and to shape global data governance? Berlin: Global Solutions Initiative; 2021 (https://www.global-solutions-initiative.org/global-table/ai-and-data-governance/#:~:text=Under%20Japan%E2%80%99s%20presidency%2C%20the%20G20%20endorsed%20Principles%20for,pursuit%20of%20beneficial%20outcomes%20for%20people%20and%20planet.%E2%80%9D, accessed April 2021).

83. OECD AI policy observatory. Paris: Organization for Economic Co-operation and Development; 2019 (https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf, accessed 5 February 2020).

84. Unboxing artificial intelligence: 10 steps to protect human rights. Strasbourg: Council of Europe; 2019 (https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64, accessed 6 February 2020).

85. Ethics guidelines for trustworthy AI. Brussels: European Commission; 2019 (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, accessed 13 February 2020).

86. AI utilisation guidelines. The Conference Towards AI Society. Paris: Organization for Economic Co-operation and Development; 2019 (https://www.oecd.ai/dashboards/policy-initiatives/2019-data-policyInitiatives-24346, accessed 5 February 2021).

87. Governance principles for the new generation artificial intelligence – Developing responsible artificial intelligence. China Daily, 17 June 2019 (http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html, accessed 20 April 2021).

88. Beijing AI principles. Beijing: Beijing Academy of Artificial Intelligence; 25 May 2019. (https://www.baai.ac.cn/news/beijing-ai-principles-en.html, accessed 20 April 2021).

89. Singapore wins international award for its artificial intelligence governance and ethics initiatives. Singapore: InfoComm Media Development Authority, 9 April 2019 (https://www.imda.gov.sg/news-and-events/Media-Room/Media-Releases/2019/singapore-wins-international-award-for-its-artificial-intelligence-governance-and-ethics-initiatives, accessed 16 February 2020).

90. Singapore Computer Society, InfoComm Media Development Authority. AI Ethics & Governance Body of Knowledge. Singapore: Singapore Computer Society; 2020 (https://ai-ethics-bok.scs.org.sg/about, accessed 9 December 2020).

91. African Union High Level Panel on Emerging Technologies (APET). Addis Ababa: African Union Development Agency; 2019 (https://www.nepad.org/microsite/african-union-high-level-panel-emerging-technologies-apet, accessed 17 January 2021).

92. Declaration of Astana. Global Conference on Primary Health Care, Astana, 25-26 October 2018. Geneva: World Health Organization; 2018 (https://www.who.int/docs/default-source/primary-health/declaration/gcphc-declaration.pdf, accessed 14 February 2020).

93. International Bioethics Committee. Report of the IBC on big data and health. Paris: United Nations Educational, Cultural and Scientific Organization; 2017 (https://unesdoc.unesco.org/ark:/48223/pf0000248724, accessed 20 February 2020).

94. World Commission on the Ethics of Scientific Knowledge and Technology. Report of COMEST on robotics ethics. Paris: United Nations Educational, Cultural and Scientific Organization; 2017 (https://unesdoc.unesco.org/ark:/48223/pf0000253952, accessed 20 February 2020).

95. Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of artificial intelligence. Paris: United Nations Educational, Cultural and Scientific Organization; 2019 (https://unesdoc.unesco.org/ark:/48223/pf0000367422, accessed 20 February 2020).

96. Guidance. Code of conduct for data-driven health and care technology. London: Department of Health and Social Care; 2019 (https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology, accessed 20 February 2020).

97. The Lancet, Financial Times Commission. Governing health futures 2030: Growing up in a digital world. Geneva: Global Health Centre, The Graduate Institute; 2021 (https://www.governinghealthfutures2030.org/#:~:text=For%20the%20first%20time%2C%20a,support%20attainment%20of%20the%20third, accessed 4 March 2021).

98. French bioethics law: an original participatory approach for the National Bioethics Consultation. Paris: Institut Pasteur, 2 September 2019 (https://www.pasteur.fr/en/home/research-journal/reports/french-bioethics-law-original-participatory-approach-national-bioethics-consultation, accessed 16 April 2021).

99. Ross WD. The right and the good. Oxford: Clarendon Press; 1930.

100. Beauchamp TL, Childress JF. The principles of biomedical ethics. 5th edition. New York City (NY): Oxford University Press; 2001.

101. Public debate. Strasbourg: Council of Europe; 2021 (https://www.coe.int/en/web/bioethics/public-debate, accessed 17 August 2020).

102. Morozov E. To save everything, click here. New York City (NY): Public Affairs; 2014.

103. Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019 (https://nam.edu/artificial-intelligence-special-publication/, accessed 18 November 2020).

104. Gasser U, Ienca M, Scheibner J, Sleigh J, Vayena E. Digital tools against COVID-19: Taxonomy, ethical challenges, and navigation aid. Lancet Digit Health. 2020;2[8]:e425-34.

105. Fenech M, Strukelj N, Buston O. The ethical, social, and political challenges of artificial intelligence in healthcare. London: Future Advocacy; 2018 (https://cms.wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf, accessed November 2020.

106. London AJ. Groundhog day for medical artificial intelligence. Hastings Centre Rep. 2018; 48[3]: doi: 10.1002/hast.842.

107. In tech-driven 21st century, achieving global development goals requires closing digital gender divide. UN News, 15 March 2019 (https://news.un.org/en/story/2019/03/1034831, accessed 16 November 2020).

108. The age of digital interdependence: Report of the United Nations Secretary-General's High-level Panel on Digital Cooperation. New York City (NY): United Nations; 2019 (https://www.un.org/en/pdfs/HLP%20on%20Digital%20Cooperation%20Report%20Executive%20Summary%20-%20ENG.pdf, accessed 14 November 2020).

109. Schwerhoff G, Sy M. Where the sun shines. Washington DC: International Monetary Fund Finance and Development; 2020 (https://www.imf.org/external/pubs/ft/fandd/2020/03/pdf/powering-Africa-with-solar-energy-sy.pdf, accessed 11 February 2021).

110. SDG7: Access to affordable, reliable, sustainable and modern energy for all. Paris: International Energy Agency; 2020 (https://www.iea.org/reports/sdg7-data-and-projections/access-to-electricity, accessed 29 November 2020).

111. Winslow J. America's digital divide. Pew Trust Magazine, 26 July 2019 (https://www.pewtrusts.org/en/trust/archive/summer-2019/americas-digital-divide, accessed 12 September 2020).

112. Buying a smartphone on the cheap? Privacy might have to be the price you have to pay. London: Privacy International; 2019 (https://privacyinternational.org/long-read/3226/buying-smart-phone-cheap-privacy-might-be-price-you-have-pay, accessed 23 February 2021).

113. Vayena E, Blassime A. Biomedical big data: New models of control over access, use, and governance. Bioethical Inquiry. 2017;14:501-13.

114. Evolving health data ecosystem. Geneva: World Health Organization; 2016 (https://www.who.int/ehealth/resources/ecosystem.pdf?ua=1, accessed 1 March 2021).

115. Vayena E, Dzenowagis J, Langfeld M. Evolving health data ecosystem. Geneva: World Health Organization; 2016 (https://www.who.int/ehealth/resources/ecosystem.pdf?ua=1, accessed 17 April 2021).

116. McNair D, Price WN. Health care AI: Law, regulation, and policy. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019.

117. Xafis V, Schaefer GO, Labude MK, Brassington I, Ballantyne A, Lim HY et al. An ethics framework for big data in health and research. Asian Bioethics Rev. 2019;11:227-54.

118. White paper: On artificial intelligence – A European approach to excellence and trust. Brussels: European Commission; 2020 (https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, accessed 8 November 2020).

119. Mozur P, Zhong R, Krolik A. In coronavirus fight, China gives citizens a color code, with red flags. The New York Times, 1 March 2020 (https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html, accessed 14 June 2020).

120. Angwadi centres, digital tracking in India's blueprint for COVID-19 vaccine drive. The Wire (Science), 6 November 2020 (https://science.thewire.in/health/anganwadi-centres-digital-tracking-in-indias-blueprint-for-covid-19-vaccination-drive/, accessed 13 January 2021).

121. Immunity passports and COVID-19: An explainer. London: Privacy International; 2020 (https://privacyinternational.org/explainer/4075/immunity-passports-and-covid-19-explainer, accessed 30 November 2020).

122. Fisher M, Han CS. How South Korea flattened the curve. The New York Times, 23 March 2020 (https://www.nytimes.com/2020/03/23/world/asia/coronavirus-south-korea-flatten-curve.html, accessed 7 December 2020).

123. The looming disaster of immunity passports and digital identity. London: Privacy International; 2020 (https://privacyinternational.org/long-read/4074/looming-disaster-immunity-passports-and-digital-identity, accessed 12 November 2020).

124. A fair shot: Ensuring universal access to COVID-19 diagnostics, treatments, and vaccines. London: Amnesty International; 2020 (https://www.amnesty.org/download/Documents/POL3034092020ENGLISH.PDF, accessed 16 December 2020).

125. Zuboff S. The age of surveillance capitalism. London: Principle Books; 2019.

126. Illmer, Andreas, Singapore reveals COVID privacy data available to police. BBC News, 5 January 2021 (https://www.bbc.com/news/world-asia-55541001, accessed 11 February 2021).

127. Chee K. Bill introduced to make clear TraceTogether, SafeEntry data can be used to look into only 7 types of serious crimes. Straits Times, 1 February 2021 (https://www.straitstimes.com/singapore/proposed-restrictions-to-safeguard-personal-contact-tracing-data-will-override-all-other, accessed 22 February 2021).

128. Price WN II, Cohen IG. Privacy in the age of medical big data. Nature Med. 2019;25[1]:37-43.

129. Copeland R. Google's Project Nightingale gathers personal health data on millions of Americans. Wall Street Journal, 11 November 2019 (https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790, accessed 12 November 2020).

130. Wood M. U Chicago Medicine collaborates with Google to use machine learning for better health care. At the Forefront: U Chicago Medicine, 17 May 2017 (https://www.uchicagomedicine.org/forefront/research-and-discoveries-articles/uchicago-medicine-collaborates-with-google-to-use-machine-learning-for-better-health-care, accessed 19 January 2021).

131. Shachar C, Gerke S, Minssen T. Is data sharing caring enough about patient privacy? Part I: The background. Cambridge (MA): Bill of Health, Harvard Law, Petrie Flom Center; 2019 (https://blog.petrieflom.law.harvard.edu/2019/07/26/is-data-sharing-caring-enough-about-patient-privacy-part-i-the-background/, accessed 13 March 2021).

132. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M et al. Scalable and accurate deep learning with electronic health records. npj Digital Med. 2018;1:18.

133. Andanda P. Ethical and legal governance of health-related research that use digital data from user-generated online health content. Inf Commun Soc. 2020;23[8]:1154-69.

134. Fussell S. Google's totally creepy, totally legal health-data harvesting. The Atlantic, 14 November 2019 (https://www.theatlantic.com/technology/archive/2019/11/google-project-nightingale-all-your-health-data/601999/, accessed 30 November 2020).

135. Lewis P, Conn D, Pegg D. UK government using confidential patient data in coronavirus response. The Guardian, 12 April 2020 (https://www.theguardian.com/world/2020/apr/12/uk-government-using-confidential-patient-data-in-coronavirus-response, accessed 30 November 2020).

136. Hern A. Anonymous browsing data can be easily exposed, researchers reveal. The Guardian, 1 August 2017 (https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers, accessed 12 February 2020).

137. Ross C. At Mayo Clinic, sharing patient data with companies fuels AI innovation – and concerns about consent. STAT News, 3 June 2020 (https://www.statnews.com/2020/06/03/mayo-clinic-patient-data-fuels-artificial-intelligence-consent-concerns/, accessed 18 November 2020).

138. Mann L. Left to other peoples' devices? A political economy perspective on the Big Data revolution in development. Dev Change. 2017;49[2]:doi: 10.1111/dech.12347.

139. Hariri Y. How to survive the 21st century. Geneva: World Economic Forum; 2020 (https://www.weforum.org/agenda/2020/01/yuval-hararis-warning-davos-speech-future-predications/, accessed 18 November 2020).

140. Krutzinna J, Taddeo M, Floridi L. Enabling posthumous medical data donation: A plea for the ethical utilisation of personal health data. In: Krutzinna J, Floridi L, editors, The ethics of medical data donation (Philosophical Studies Series, Vol 137). Cham: Springer; 2019.

141. Shaw DM, Gross JV, Erren TC. Why you should donate your health data (as well as your organs) when you die. STAT News, 14 February 2017. (https://www.statnews.com/2017/02/14/donate-health-data-death/, accessed 18 September 2020).

142. General Data Protection Regulation. Article 27. Brussels: European Union; 2016 (https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed 18 March 2021).

143. Malgieri G. RIP: Rest in privacy or rest in (quasi-)property? Personal data protection of deceased data subjects between theoretical scenarios and national solutions. In: Leenes R, van Brackel R, Gutwirth S, De Hert P, editors. Data protection and privacy: The Internet of bodies. Brussels: Hart; 2018) (https://ssrn.com/abstract=3185249, accessed 12 February 2021).

144. At a glance: De-identification, anonymisation, and pseudo-anonymisation under the GDPR. Boulder (CO): Bryan Cave Leighton Paisner; 2017 (https://www.bclplaw.com/en-US/insights/at-a-glance-de-identification-anonymization-and-pseudonymization-1.html, accessed 12 September 2020).

145. General Data Protection Regulation, Article 5. Brussels: European Union; 2016 (https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed 18 March 2021).

146. Bari L, O'Neill D. Rethinking patient data privacy in the era of digital health. Health Affairs, 12 December 2019 (https://www.healthaffairs.org/do/10.1377/hblog20191210.216658/full/, accessed 23 November 2020).

147. Rocher L, Hendrickx JM, de Montjoye Y. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10:3069.

148. May T. Sociogenetic risks – ancestry DNA testing, third-party identity, and protection of privacy. NEJM. 2018;379:410-2.

149. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46[3]:205-11.

150. Yeung K. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Strasbourg: Council of Europe; 2019 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3286027, accessed 15 March 2020).

151. Habli I, Lawton T, Porter Z. Artificial intelligence in healthcare: Accountability and safety. Bull World Health Organ. 2020;98:251-6.

152. Hurtgen H, Kerkhoff S, Lubatschowski J, Möller M. Rethinking AI talent strategy as automated machine learning comes of age. New York City (NY): McKinsey and Co., 2020 (https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/rethinking-ai-talent-strategy-as-automated-machine-learning-comes-of-age, accessed 4 November 2020).

153. Dixon-Woods M, Pronovost PJ. Patient safety and the problem of many hands. BMJ Qual Saf. 2016;25[7]:485-8.

154. Van de Poel I, Royakkers L, Zwart SD, de Lima T, Doorn N, Fahlquist JN. Moral responsibility and the problem of many hands. New York City (NY): Routledge; 2015.

155. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2020;doi:10.1136/medethics-2019-105860.

156. Metcalf J, Moss E, Boyd D. Owning ethics: Corporate logics, Silicon Valley, and the institutionalisation of ethics. Soc Res. 2019;82[2]:449-76.

157. Whitaker M, Crawford K, Dobbe R, Fried G, Kaziunas E, Mathur V et al. AI Now report 2018. New York City (NY): AI Now Institute; 2018 (https://ainowinstitute.org/AI_Now_2018_Report.pdf, accessed 13 December 2020).

158. Vincent J. The problem with ethics: Is Big Tech's embrace of AI ethics boards actually helping anyone. The Verge, 3 April 2019 (https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech, accessed 14 January 2021).

159. Artificial intelligence in healthcare. London: Academy of Medical Royal Colleges; 2019 (https://www.aomrc.org.uk/reports-guidance/artificial-intelligence-in-healthcare/, accessed 18 July 2020).

160. In brief: Artificial intelligence in health care. Stockholm: Swedish National Council on Medical Ethics; 2020 (https://smer.se/en/publications/?date=2020-5, accessed 12 July 2020).

161. Duran JM. Computer simulations in science and engineering. Cham: Springer; 2018 (https://link.springer.com/book/10.1007%2F978-3-319-90882-3, accessed April 2021).

162. Humphreys P. The philosophical novelty of computer simulation methods. Synthese. 2009;169[3]:615-26.

163. Topol E. Twitter, 7 January 2019 (https://twitter.com/EricTopol/status/1082363519675248640/photo/1, accessed 15 March 2021).

164. Pasquale F. The black box society: The secret algorithms that control money and Information. Camridge (MA): Harvard University Press; 2015.

165. London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Rep. 2019;49[1]:15-21.

166. Durán JM, Formanek N. Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds Machines. 2018;28:645-66.

167. Cohen IG. Informed consent and medical artificial intelligence: What to tell the patient? Georgetown Law J. 2020;108:1425.

168. Minssen T, Rajam N, Bogers M. Clinical trial data transparency and GDPR compliance: Implications for data sharing and open innovation. Sci Public Policy. scaa014: doi.org/10.1093/scipol/scaa014.

169. Mullin E. Healthcare is the next battleground for Big Tech. One Zero, 27 January 2020 (https://onezero.medium.com/health-care-is-the-next-battleground-for-big-tech-477a7263974, accessed 14 January 2021).

170. Turea M. How the big 4 tech companies are leading innovation. Healthcare Weekly, 27 February 2019 (https://healthcareweekly.com/how-the-big-4-tech-companies-are-leading-healthcare-innovation/, accessed 15 January 2021).

171. A look back at Alphabet's moves in 2019. MobiHealthNews, 13 December 2019 (https://healthcareweekly.com/how-the-big-4-tech-companies-are-leading-healthcare-innovation/, accessed 12 November 2020).

172. Our work with Google Health UK. London: NHS Royal Free London; 2019 (https://www.royalfree.nhs.uk/patients-visitors/how-we-use-patient-information/our-work-with-deepmind/, accessed 12 February 2021).

173. Shepherd C. China's online health platforms boom in wake of coronavirus. The Financial Times, 16 December 2020 (https://www.ft.com/content/22b22543-0fb5-4a8a-8ec0-e3fd067a5190, accessed 1 February 2021).

174. Bridging gaps in healthcare industry with technology. Shenzhen: Tencent Holdings Ltd; 2019 (https://www.tencent.com/en-us/articles/2200933.html, accessed 6 November 2020).

175. How Baidu, Alibaba, and Tencent aim to disrupt Chinese health care. Forkast, 28 January 2020 (https://forkast.news/baidu-alibaba-tencent-china-health-care-blo/, accessed 13 February 2021).

176. Jourdan A. AI ambulances and robot doctors: China seeks digital salve to ease hospital strain. Reuters, 28 June 2018 (https://de.reuters.com/article/us-china-healthcare-tech-idUKKBN1JO1VB, accessed 24 November 2020).

177. Goodman K. Ethics, medicines, and information technology: Intelligent machines and the transformation of health care. Cambridge: Cambridge University Press; 2016.

178. Chen C. Only seven of Stanford's first 5000 vaccines were designated for medical residents. ProPublica, 18 December 2020 (https://www.propublica.org/article/only-seven-of-stanfords-first-5-000-vaccines-were-designated-for-medical-residents, accessed 18 February 2021).

179. Hariri YN. Homo deus: A brief history of tomorrow. London: Vintage; 2015.

180. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology. 2019;290[2]:456-64.

181. Chen JH, Beam A, Saria S, Mendonça E. Potential trade-offs and unintended consequences of AI. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019.

182. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572:116-9.

183. Morley J, Caio C, Machado V, Burr C, Cowls J, Joshi I et al. The debate on the ethics of AI in health care: a reconstruction and critical review. Oxford: Oxford Internet Institute; 2019 (https://digitalethicslab.oii.ox.ac.uk/wp-content/uploads/sites/87/2019/11/The-Debate-on-the-EThics-of-AI-in-Health-Care-pre-print-.pdf, accessed 15 July 2020).

184. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 2015;26[1]:15030.

185. Marcus J, Hurley L, Krakower D, Alexeeff S, Silverberg M, Volk J. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. Lancet HIV. 2019;6:10.1016/S2352-3018[19]30137-7.

186. Urtubey y una insólita propuesta de "prever" embarazos adolescents [Urtubey and an unusual proposal to "anticipate" adolescent pregnancies]. Diario de Cuyo, 11 April 2018 (https://www.diariodecuyo.com.ar/argentina/Urtubey-y-una-insolita-propuesta-de-prever-embarazos-adolescentes-20180411-0081.html, accessed 12 February 2021).

187. Peña P, Varon J. Decolonising AI: A transfeminist approach to data and social justice. In: Artificial intelligence: Human rights, social justice and development. Global Information Society Watch. Association for Progressive Communications; 2019 (https://giswatch.org/sites/default/files/gisw2019_web_th4.pdf, accessed 12 February 2021).

188. Sobre la predicción automática de embarazos adolescents [On automatic prediction of adolescent pregnancies]. Buenos Aires: Universidad de Buenos Aires, Laboratorio de Inteligencia Artificial Aplicada; 2018 (https://www.dropbox.com/s/r7w4hln3p5xum3v/%5BLIAA%5D%20Sobre%20la%20predicci%C3%B3n%20autom%C3%A1tica%20de%20embarazos%20adolescentes.pdf, accessed 12 February 2021).

189. Venturini J. Surveillance and social control: How technology reinforces structural inequality in Latin America. London: Privacy International; 2019 (https://privacyinternational.org/news-analysis/3263/surveillance-and-social-control-how-technology-reinforces-structural-inequality, accessed 12 February 2021).

190. Ortiz Freuler J, Iglesias C. Algorithms and artificial intelligence in Latin America: A study of implementation by governments in Argentina and Uruguay. Washington DC: World Wide Web Foundation; 2018 (http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf, accessed 12 February 2021).

191. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366[6464]:447-53.

192. Price WN II. Medical AI and contextual bias (U of Michigan Public Law Research Paper No. 632). Harvard J. Law Technol. 2019;66 (https://ssrn.com/abstract=3347890, accessed 13 September 2022).

193. Minssen T, Gerke S, Aboy M, Price N, Cohen G. Regulatory responses to medical machine learning J Law Biosci. lsaa002 (https://doi.org/10.1093/jlb/lsaa002, accessed 13 September 2022).

194. Gerke S, Minssen T, Yu H, Cohen IG. Ethical and legal issues of ingestible electronic sensors. Nat Electron. 2019;2:329-34.

195. Simonite T. How an algorithm blocked kidney transplants to patients. Wired Magazine, 26 October 2020 (https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/, accessed 12 November 2020).

196. Benjamin R. Assessing risk, automating racism. Science. 2019;366[6464]:421-2.

197. Lashbrook A. AI-driven dermatology could leave dark-skinned patients behind. The Atlantic, 16 August 2018 (https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/, accessed 14 December 2020).

198. Bridging the digital gender divide: Include, upskill, innovate. Paris: Organization for Economic Co-operation and Development; 2018 (http://www.oecd.org/digital/bridging-the-digital-gender-divide.pdf, accessed 3 November 2020).

199. Munshi N. How unlocking the secrets of African DNA could change the world. The Financial Times, 5 March 2020 (https://www.ft.com/content/eed0555c-5e2b-11ea-b0ab-339c2307bcd4, accessed 2 November 2020).

200. Devlin H., Genetics research "biased towards studying white Europeans". The Guardian, 8 October 2018 (https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towards-studying-white-europeans, accessed 2 November 2020).

201. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digital Med. 2020;3:81.

202. Rose E. How self-tracking apps exclude women. The Atlantic, 15 December 2014 (https://www.theatlantic.com/technology/archive/2014/12/how-self-tracking-apps-exclude-women/383673/, accessed 30 January 2021).

203. Hern A. Fault in NHS Covid app meant thousands at risk did not quarantine. The Guardian, 2 November 2020 (https://www.theguardian.com/world/2020/nov/02/fault-in-nhs-covid-app-meant-thousands-at-risk-did-not-quarantine, 20 January 2020).

204. Contag M, Li G, Pawlowski A, Domke F, Levchenko K, Holz T et al. How they did it: An analysis of emission defeat devices in modern automobiles. In: IEEE Symposium on Security and Privacy. San Diego (CA): University of California at San Diego; 2017 (https://cseweb.ucsd.edu/~klevchen/diesel-sp17.pdf, accessed 8 November 2020).

205. Baraniuk C. How tech bugs could be killing thousands in our hospitals. New Scientist, 16 May 2018 (https://www.newscientist.com/article/mg23831781-700-how-tech-bugs-could-be-killing-thousands-in-our-hospitals/, accessed 20 August 2020).

206. Xu K, Soucat A, Kutzin J, Siroka A, Aranguren Garcia M, Dupuy J et al. Global spending on health: A world in transition. Geneva; World Health Organization; 2019 (https://apps.who.int/iris/bitstream/handle/10665/330357/WHO-HIS-HGF-HF-WorkingPaper-19.4-eng.pdf?ua=1, accessed 25 February 2021).

207. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly. 2018;148:w14571.

208. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: The privacy bounds of human mobility. Sci Rep. 2013;3:1376.

209. Telemedicine: Opportunities and developments in Member States (Global Observatory for eHealth Series, Vol. 2). Geneva: World Health Organization; 2010 (https://www.who.int/goe/publications/goe_telemedicine_2010.pdf, accessed 17 February 2021).

210. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Health J. 2019;6[2]:94-8.

211. Health workforce. Geneva: World Health Organization; 2021 (https://www.who.int/health-topics/health-workforce#tab=tab_1, accessed 7 December 2020).

212. Parikh RK. Should doctors play along with the uberization of health care. Slate, 14 June 2017 (https://slate.com/technology/2017/06/should-doctors-play-along-with-the-uberization-of-health-care.html, accessed 11 November 2020).

213. Gawande A. Why doctors hate their computers. The New Yorker, 5 November 2018 (https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers, accessed 12 December 2020).

214. COVID-19 response: Corporate exploitation. London: Privacy International; 2020 (https://privacyinternational.org/news-analysis/3592/covid-19-response-corporate-exploitation, accessed 13 February 2021).

215. Powles J, Hodson H. Google Deepmind and healthcare in an age of algorithms. Health Technol (Berl). 2017;7[4]:351-67.

216. Ballantyne A, Stewart C. Big data and public-private partnerships on healthcare and research. Asian Bioethics Rev. 2019;11:315-26.

217. Hodson H. Revealed: Google AI has access to huge haul of NHS patient data. New Scientist, 29 April 2016 (https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/, accessed 29 November 2020).

218. Durkee A. Facebook to pay millions for allegedly mishandling user data (again). Vanity Fair, 30 January 2020 (https://www.vanityfair.com/news/2020/01/facebook-settlement-facial-recognition-illinois-privacy, accessed 29 November 2020).

219. Mergers: Commission clears acquisition of Fitbit by Google, subject to conditions. Brussels: European Commission; 2020. (https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2484, accessed 12 February 2021).

220. Mergers: Commission opens in-depth investigation into the proposed acquisition of Fitbit by Google. Brussels: European Commission; 2020 (https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1446, accessed 12 November 2020).

221. Competition and data. London: Privacy International; 2021 (https://privacyinternational.org/learn/competition-and-data, accessed 6 February 2021).

222. Bridging gaps in healthcare industry with technology. Shenzhen: Tencent Holdings Ltd; 2019 (https://www.tencent.com/en-us/articles/2200933.html, accessed 24 November 2020).

223. Veale M. Privacy is not the problem with the Google-Apple contact-tracing toolkit. The Guardian, 1 July 2020 (https://www.theguardian.com/commentisfree/2020/jul/01/apple-google-contact-tracing-app-tech-giant-digital-rights, accessed 24 November 2020).

224. Hao K. Training an AI model can emit as much carbon as five cars in their lifetime. MIT Technology Review, 6 June 2019 (https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/, accessed 12 December 2020).

225. DeWeerdt S. It's time to talk about the carbon footprint of artificial intelligence. Anthropocene, 10 November 2020 (https://www.anthropocenemagazine.org/2020/11/time-to-talk-about-carbon-footprint-artificial-intelligence/, accessed 27 February 2021).

226. Climate change. Geneva: World Health Organization; 2021 (https://www.who.int/health-topics/climate-change#tab=tab_1, accessed 27 February 2021).

227. Hao K. We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review, 4 December 2020 (https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/?utm_source=Nature+Briefing&utm_campaign=ebee85e120-briefing-dy-20201208&utm_medium=email&utm_term=0_c9dfd39373-ebee85e120-44944633, accessed 12 December 2020).

228. van den Hoven J, Vermaas PE, van de Poel I, editors. Handbook of ethics, values, and technological design: Sources, theories, values, and application domains. Cham: Springer; 2015 (https://www.springer.com/gp/book/9789400769694#aboutAuthors, accessed April 2021).

229. Aizenberg E, van den Hoven J. Designing for human rights in AI. Big Data Society. 2020;July-December:1-14.

230. Statement regarding the ethical implementation of artificial intelligence systems (AIS) for addressing the COVID-19 pandemic. New York City (NY): Institute of Electrical and Electronic Engineering; 2020 (https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/gieais-covid.pdf, accessed 13 September 2022).

231. Independent High-level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Brussels: European Commission; 2019 (https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top, accessed 12 November 2020).

232. Whitaker K. Citizen science platform with Autistica. London: The Alan Turing Institute; 2019 (https://www.turing.ac.uk/research/research-projects/citizen-science-platform-autistica, accessed 12 November 2020).

233. The elements of informed consent: A toolkit. V.3. Seattle (WA): Sage Bionetworks; 2020 (https://sagebionetworks.org/wp-content/uploads/2020/01/SageBio_EIC-Toolkit_V3_21Jan20_final.pdf, accessed 13 November 2020).

234. EPI-Brain webpage. Geneva: World Health Organization; 2020 (https://www.epi-brain.com/, accessed 3 November 2020).

235. IEEE P7000. IEEE draft model process for addressing ethical concerns during system design. Piscataway (NJ): IEEE Standards Association; 2016 (https://standards.ieee.org/project/7000.html, accessed April 2021).

236. Understanding patient data. London: Wellcome Trust; 2019 (https://understandingpatientdata.org.uk/, accessed 13 February 2020).

237. Sharing anonymised patient-level data where there is a mixed public and private benefit – a new report. London: Health Research Authority; 2019 (https://www.hra.nhs.uk/about-us/news-updates/sharing-anonymised-patient-level-data-where-there-mixed-public-and-private-benefit-new-report/, accessed 17 February 2020).

238. Artificial intelligence and health, summary report of a roundtable held on 16 January 2019. London: Academy of Medical Sciences; 2019 (https://acmedsci.ac.uk/file-download/77652269, accessed 17 February 2020).

239. Our data driven future in healthcare. People and partnerships at the heart of health-related technologies. London: Academy of Medical Sciences; 2018 (https://acmedsci.ac.uk/file-download/74634438, accessed 17 February 2020).

240. Trust in technology. London: HSBC Holdings Ltd; undated (http://www.hsbc.com/trust-in-technology-report, accessed April 2021).

241. Trustworthy AI in health: Background paper for the G20 AI dialogue, digital economy, and trade. Paris: Organization for Economic Co-operation and development; 2020 (https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf, accessed 11 November 2020).

242. Blog: ICO regulatory sandbox. London: Information Commissioner's Office; 2020 (https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/11/sandbox-helps-develop-innovative-tools-to-combat-financial-crime/, accessed 13 February 2021).

243. Fihn SD, Saria S, Mendonça E, Hain S, Matheny M, Shah N et al. Deploying AI in clinical settings. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019.

244. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ. 2019;5[2]:e16048.

245. What is impact assessment. Fact sheet. Fargo (ND): International Association for Impact Assessment; 2009 (https://www.iaia.org/uploads/pdf/What_is_IA_web.pdf, accessed 9 November 2020).

246. Guiding principles on business and human rights: Implementing the United Nations protect, respect, and remedy framework. Geneva: Office of the High Commissioner of Human Rights; 2011 (https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf, accessed 9 November 2020).

247. Human rights impact assessments. National Action Plans on Business and Human Rights. Copenhagen: Danish Institute for Human Rights; 2020 (https://globalnaps.org/issue/human-rights-impact-assessments/, accessed 19 November 2020).

248. French corporate duty of vigilance law. Brussels: European Coalition of Corporate Justice; 2017 (https://corporatejustice.org/documents/publications/french-corporate-duty-of-vigilance-law-faq.pdf, accessed 19 November 2020).

249. Marlow J. New EU law requiring human rights due diligence on the cards for 2021. Blog, 28 July 2020. Paris: Linklaters LLP (https://www.linklaters.com/en/insights/blogs/linkingesg/2020/july/new-eu-law-requiring-human-rights-due-diligence-on-the-cards-for-2021, accessed 19 November 2020).

250. Algorithmic impact assessments: A practical framework for public agency accountability. New York City (NY): AI Now Institute; 2018 (https://www.ftc.gov/system/files/documents/public_comments/2018/08/ftc-2018-0048-d-0044-155168.pdf, accessed 19 November 2020).

251. MacCarthy M. An examination of the Algorithmic Accountability Act of 2019. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Amsterdam: Institute for Information Law; 2019 (https://www.ivir.nl/publicaties/download/Algorithmic_Accountability_Oct_2019.pdf, accessed 19 November 2020).

252. Data protection impact assessment (DPIA). How to conduct a data protection impact assessment (template included). GPPR.EU (https://gdpr.eu/data-protection-impact-assessment-template/, accessed 19 November 2020).

253. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C et al. Machine behaviour. Nature. 2019;568:477-86 [2019].

254. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA. 2019;322[18]:1765-6.

255. Price WN. Artificial intelligence in healthcare: Applications and legal implications. The SciTech Lawyer. 2017;14[1]. University of Michigan Law School Scholarship Repository (https://repository.law.umich.edu/cgi/viewcontent.cgi?article=2932&context=articles, accessed April 2021).

256. Gerke S, Minssen T, Cohen G. Chapter 12. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Cambridge (MA): Academic Press;2020:295-336.

257. Ordish J. Briefing. Legal liability for machine learning in healthcare. Cambridge: PHG Foundation; 2018 (https://www.phgfoundation.org/documents/briefing-note-legal-liability-for-machine-learning-in-healthcare.pdf, accessed 22 November 2020).

258. Minssen T, Mimler M, Mak V. When does stand-alone software qualify as a medical device in the European Union? The Cjeu's decision in Snitem and what it implies for the next generation of medical devices. Med Law Rev. 2020;28[3]:615-24.

259. Evans BJ, Pasquale FA. Product liability suits for FDA-regulated AI/ML software (Brooklyn Law School, Legal Studies Paper No. 656). In: Cohen IG, Minssen T, Price WN II, Robertson C, Shachar C, editors. The future of medical device regulation: Innovation and protection. Cambridge: Cambridge University Press; 2021 (https://ssrn.com/abstract=3719407, accessed April 2021).

260. Report from the Commission to the European Parliament, the Council, and the European Economic and Social Committee: Report on the safety and liability implications of artificial intelligence, the Internet of Things and robotics. Brussels: European Commission; 2020 (https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064, accessed 15 July 2020).

261. Price WN II. Medical malpractice and black-box medicine (University of Michigan Public Law Research Paper No. 536). In: Cohen IG, Minssen T, Price WN II, Robertson C, Shachar C, editors. The future of medical device regulation: Innovation and protection. Cambridge: Cambridge University Press; 2021 (https://ssrn.com/abstract=2910417, accessed April 2021).

262. Husgen J. Product liability suits involving drug or device manufacturers and physicians: the learned intermediary doctrine and the physician's duty to warn. MO Med. 2014;111[6]:478-81.

263. Thomas S. Artificial intelligence and medical liability (Part II). Bill of Health. 10 February 2017. (https://blog.petrieflom.law.harvard.edu/2017/02/10/artificial-intelligence-and-medical-liability-part-ii/, accessed 17 November 2020).

264. No fault compensation in New Zealand: Harmonizing injury compensation, provider accountability, and patient safety. Commonwealth Fund, 24 February 2006 (https://www.commonwealthfund.org/publications/journal-article/2006/feb/no-fault-compensation-new-zealand-harmonizing-injury, accessed 20 March 2021).

265. McNair D, Price WN. Health care AI: Law, regulation, and policy. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019.

266. Digital health (A/71/A/CONF./1). Seventy-first World Health Assembly. Geneva: World Health Organization; 2018 (https://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_ACONF1-en.pdf, accessed 20 November 2020).

267. Elements of informed consent. Seattle (WA): Sage Bionetworks; 2020 (https://sagebionetworks.org/tools_resources/elements-of-informed-consent/, accessed 14 March 2021).

268. Cohen IG. Is there a duty to share healthcare data? In: Cohen IG, Lynch HF, Vayena E, Gasser U, editors. Big data, health law, and bioethics. Cambridge: Cambridge University Press, 2018:209-22.

269. Otake T. Medical big data to be pooled for disease research and drug development in Japan. Japan Times, 15 May 2017 (https://www.japantimes.co.jp/news/2017/05/15/reference/medical-big-data-pooled-disease-research-drug-development-japan/#:~:text=The%20law%2C%20commonly%20called%20Jisedai,the%20development%20of%20new%20drugs, accessed 14 February 2021).

270. Regulations on data governance – questions and answers. Brussels: European Commission; 2020 (https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2103, accessed 13 December 2020).

271. Spector-Bagdady K, Hutchinson R, O'Brien Kaleba E, Kheterpal S. Sharing health data and biospecimens with industry – A principle-driven, practical approach. NEJM. 2020;382[22]:2072-5.

272. Greenberg Z. What is the blood of a poor person worth? New York Times, 1 February 2019 (https://www.nytimes.com/2019/02/01/sunday-review/blood-plasma-industry.html, accessed 17 November 2020).

273. Data protection guide. London: Privacy International; 2018 (https://privacyinternational.org/report/2255/data-protection-guide-complete, accessed 14 February 2021).

274. Bowan N. After seven year wait, South Africa's data protection law enters into force. Portsmouth (NH): International Association of Privacy Professionals; 2020 (https://iapp.org/news/a/after-a-7-year-wait-south-africas-data-protection-act-enters-into-force/, accessed 15 February 2021).

275. National Data Guardian: What we do. London: HM Government; 2020 (https://www.gov.uk/government/organisations/national-data-guardian/about, accessed 5 November 2020).

276. Our charter: Tūtohinga. Aukland: Te Mana Raraunga (Maori Data Sovereignty Network); 2020 (https://www.temanararaunga.maori.nz/tutohinga, accessed 5 November 2020).

277. Schnarch B. Ownership, access, control, and possession (OCAP) or self-determination applied to research. Int J Indigenous Health. 2004;1[1] (https://jps.library.utoronto.ca/index.php/ijih/article/view/28934, accessed 19 November 2020).

278. Sharing sensitive health data in a federated data consortium model. An eight-step guide. Insight report. Geneva: World Economic Forum; 2020 (http://www3.weforum.org/docs/WEF_Sharing_Sensitive_Health_Data_2020.pdf, accessed 19 November 2020).

279. Creating the right framework to realise the benefits for patients and the NHS where data underpins innovation. London: Department of Health and Social Care; 2019 (https://www.gov.uk/government/publications/creating-the-right-framework-to-realise-the-benefits-of-health-data/creating-the-right-framework-to-realise-the-benefits-for-patients-and-the-nhs-where-data-underpins-innovation, accessed 19 November 2020.

280. Bhunia P. Data futures partnership in New Zealand issues guidelines for organisations to develop social license for data use. Open Government, 27 October 2017 (https://opengovasia.com/data-futures-partnership-in-new-zealand-issues-guidelines-for-organisations-to-develop-social-license-for-data-use/#:~:text=The%20Data%20Futures%20Partnership%20is,control%20data%2Dsharing%20ecosystem.%E2%80%9D, accessed 19 November 2020).

281. WHO data principles. Geneva: World Health Organization; 2021. (https://www.who.int/data/principles, accessed 13 March 2021).

282. WHO data sharing policy: Implementation suggestions. Geneva: World Health Organization; 2020 (https://cdn-auth-cms.who.int/media/docs/default-source/world-health-data-platform/who-data-sharing-policy-implementation-suggestions-10-august-2020.pdf?sfvrsn=fd365554_2, accessed 19 April 2021).

283. Genomic data sharing policy. Bethesda (MD): National Institutes of Health; 2014 (https://www.federalregister.gov/documents/2014/08/28/2014-20385/final-nih-genomic-data-sharing-policy, accessed 12 September 2020).

284. Majumder MA, Guerrini CJ, Bollinger JM, Deegan RC, McGuire AL. Sharing data under the 21st Century Cures Act. Genet Med. 2017;19[12]:1289-94.

285. HHS finalizes historic rules to provide patients with more control of their patient data. Washington DC: Department of Health and Human Services; 2020 (https://www.hhs.gov/about/news/2020/03/09/hhs-finalizes-historic-rules-to-provide-patients-more-control-of-their-health-data.html, accessed 12 September 2020).

286. All of Us Research Program. Bethesda (MD): National Institutes of Health; 2020 (https://allofus.nih.gov/, accessed 14 November 2020).

287. European health data space. Brussels; European Commission; 2020 (https://ec.europa.eu/health/ehealth/dataspace_en, accessed 14 November 2020).

288. Digital innovation hub programme prospectus. Appendix: Principles for participation. London: Health Data Research UK; 2019 (https://www.hdruk.ac.uk/wp-content/uploads/2019/07/Digital-Innovation-Hub-Programme-Prospectus-Appendix-Principles-for-Participation.pdf, accessed 15 November 2020).

289. Ornstein C, Thomas K. Sloan Kettering's cozy deal with start-up ignites uproar. New York Times, 20 September 2018 (https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html, accessed 19 November 2020).

290. The world's most valuable resource is no longer oil, but data. The Economist, 6 May 2017 (https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource, accessed 22 August 2020).

291. Rajan A. Data is not the new oil. BBC News Online, 9 October 2017 (https://www.bbc.com/news/entertainment-arts-41559076, accessed 13 September 2020).

292. Marr B. Here's why data is not the new oil. Forbes Magazine, 5 March 2018 (https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#6a65a1453aa9, accessed 13 September 2020).

293. Hilty R. Big data: Ownership and use in the digital age. In: Seuba X, Geiger C, Penin J, editors. Intellectual property and digital trade in the age of artificial intelligence and big data (Global perspectives and challenges for the intellectual property system. Issue No. 5). Geneva: International Centre for Trade and Sustainable Development; Strasbourg: Center for International Intellectual Property Studies; 2018 (http://www.ceipi.edu/fileadmin/upload/DUN/CEIPI/Documents/Publications_CEIPI___ICTSD/CEIPI-ICTSD_Issue_5_Final.pdf, accessed 24 August 2020).

294. Minssen T, Pierce J. Big data and intellectual property in the health and life sciences. In: Cohen IG, Lynch HF, Vayena E, Gasser U, editors. Big data, health law, and bioethics. Cambridge: Cambridge University Press, 2018.

295. Andanda P. Towards a paradigm shift in governing data access and related intellectual property rights in big data and health-related research. Int Revf Intellectual Property Competition Law. 2019;50:1052-81.

296. Sherkow JS, Minssen T. AIRR data under the EU Trade Secrets Directive – Aligning scientific practices with commercial realities. In: Schovsbo J, Riis T, Minssen T, editors. The harmonization and protection of trade secrets in the EU – An appraisal of the EU Directive. Cheltenham: Edward Elgar Publishing; 2020:239-68.

297. Minssen T, Schovsbo J. Big data in the health and life sciences: What are the challenges for European competition law and where can they be found? In: Seuba X, Geiger C, Penin J, editors. Intellectual property and digital trade in the age of artificial intelligence and big data (Global perspectives and challenges for the intellectual property system. Issue No. 5). Geneva: International Centre for Trade and Sustainable Development; Strasbourg: Center for International Intellectual Property Studies; 2018 (http://www.ceipi.edu/fileadmin/upload/DUN/CEIPI/Documents/Publications_CEIPI___ICTSD/CEIPI-ICTSD_Issue_5_Final.pdf, accessed 22 August 2020).

298. European Open Science Cloud (https://www.eosc-portal.eu/, accessed 13 September 2022).

299. Corrales Compagnucci, M, Minssen T, Seitz C, Aboy M. Lost on the high seas without a safe harbor or a shield? Navigating cross-border data transfers in the pharmaceutical sector after Schrems II invalidation of the EU-US privacy shield. Eur Pharmaceut Law Rev. 2020;4[3]:153-60.

300. Abbott R. Inventive machines: Rethinking invention and patentability. In: Seuba X, Geiger C, Penin J, editors. Intellectual property and digital trade in the age of artificial intelligence and big data (Global perspectives and challenges for the intellectual property system. Issue No. 5). Geneva: International Centre for Trade and Sustainable Development; Strasbourg: Center for International Intellectual Property Studies; 2018 (http://www.ceipi.edu/fileadmin/upload/DUN/CEIPI/Documents/Publications_CEIPI___ICTSD/CEIPI-ICTSD_Issue_5_Final.pdf, accessed 22 August 2020).

301. EPO publishes grounds for its decision to refuse two patent applications naming a machine as an inventor. Munich: European Patent Office, 28 January 2020 (https://www.epo.org/news-events/news/2020/20200128.html, accessed 21 March 2021).

302. Porter J. US Patent Office rules that artificial intelligence cannot be a legal inventor. The Verge, 29 April 2020 (https://www.theverge.com/2020/4/29/21241251/artificial-intelligence-inventor-united-states-patent-trademark-office-intellectual-property, accessed 22 August 2020).

303. Aboy M, Liddell K, Crespo C, Cohen IG, Liddicoat J, Gerke S et al. How does emerging patent case law in the US and Europe affect precision medicine? Nature Biotechnol. 2019;37:1118-26.

304. West DM. The role of corporations in addressing AI's ethical dilemmas. Washington DC: Brookings Institute; 2018 (https://www.brookings.edu/research/how-to-address-ai-ethical-dilemmas/, accessed 24 August 2020).

305. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 2019;1:501-7.

306. Metz C, Wakabayashi D. Google researcher said she was fired over paper highlighting bias in AI. The New York Times, 3 December 2020 (https://nyti.ms/2I8oves, accessed 16 December 2020).

307. Cath C. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Phil Trans R Soc A. 2018;376:20180080.

308. WHO launches a chatbot on Facebook messenger to combat misinformation. Geneva: World Health Organization; 2020 (https://www.who.int/news-room/feature-stories/detail/who-launches-a-chatbot-powered-facebook-messenger-to-combat-covid-19-misinformation, accessed 29 September 2020).

309. Facebook's algorithm: A major threat to public health. Avaaz, 19 August 2020 (https://avaazimages.avaaz.org/facebook_threat_health.pdf, accessed 19 November 2020.

310. Lee D, Murphy H. Facebook accused of failing to tackle medical hoaxes. Financial Times, 20 August 2020 (https://www.ft.com/content/f33f7d61-a8df-40b9-82a8-75f2a41210bc, accessed 24 August 2020).

311. Jin KX. Keeping people safe and informed about the coronavirus. Facebook, 3 December 2020 (https://about.fb.com/news/2020/12/coronavirus/?utm_source=STAT+Newsletters&utm_campaign=01f1d5a35b-health_tech_COPY_01&utm_medium=email&utm_term=0_8cab1d7961-01f1d5a35b-151577169, accessed 16 December 2020).

312. Brodwin E. Facebook's Covid-19 misinformation campaign is based on research. The authors worry Facebook missed the message. STAT News, 1 May 2020 (https://www.statnews.com/2020/05/01/facebooks-covid-19-misinformation-campaign-is-based-on-research-the-authors-worry-facebook-missed-the-message/, accessed 13 September 2022).

313. Isaac M, Wakabayashi D, Cave D, Lee E. Facebook blocks news in Australia, diverging with Google on proposed law. The New York Times, 17 February 2021. (https://www.nytimes.com/2021/02/17/technology/facebook-google-australia-news.html, accessed 24 February 2021).

314. Taylor J. Facebook's botched Australia new ban hits health departments, charities, and its own pages. The Guardian, 18 February 2021 (https://www.theguardian.com/technology/2021/feb/18/facebook-blocks-health-departments-charities-and-its-own-pages-in-botched-australia-news-ban, accessed 24 February 2021).

315. Taylor J, McGowan M, Bland A. Misinformation runs rampant as Facebook says it may take a week before it unblocks some pages. The Guardian, 19 February 2021 (https://www.theguardian.com/technology/2021/feb/19/misinformation-runs-rampant-as-facebook-says-it-may-take-a-week-before-it-unblocks-some-pages, accessed 24 February 2021).

316. International Organization for Standardization. Geneva (https://www.iso.org/home.html, accessed 13 September 2022).

317. Health level 7 International. Ann Arbor (MI) (http://www.hl7.org/, accessed 13 September 2022).

318. Brown KV. Alphabet's Verily plans to use big data as health insurance tool. Employee Benefit News, 25 August 2020 (https://www.benefitnews.com/articles/alphabets-verily-plans-to-use-big-data-as-health-insurance-tool, accessed 12 September 2020).

319. Ackroyd AT. Tencent-backed WeDoctor makes IPO appointment in Hong Kong and writes prescription for digital healthcare post-pandemic. South China Morning Post, 4 June 2020 (https://www.scmp.com/business/banking-finance/article/3087385/tencent-backed-wedoctor-makes-ipo-appointment-hong-kong, accessed 28 August 2020).

320. Hello world: Artificial intelligence and its use in the public sector. Paris: Organization for Economic Co-operation and Development; 2019 (https://oecd-opsi.org/wp-content/uploads/2019/11/AI-Report-Online.pdf, accessed 28 August 2020).

321. The beginning of AI revolution in UAE healthcare. Global Business Outlook, 8 October 2020 (https://www.globalbusinessoutlook.com/the-beginning-of-ai-revolution-in-uae-healthcare/, accessed 5 December 2020).

322. Working document: Enforcement mechanisms for responsible #AIforAll. New Delhi: NITI Aayog; 2020 (https://niti.gov.in/sites/default/files/2020-11/Towards-Responsible-AI-Enforcement-of-Principles.pdf, accessed 12 December 2020).

323. Assessing if artificial intelligence is the right solution. London: HM Government; 2019 (https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution, accessed 28 August 2020).

324. Committee on Standards in Public Life. Artificial intelligence and public standards: report. London: HM Government; 2020 (https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report, accessed 12 February 2020).

325. Martinho-Truswell E. How AI could help the public sector, Harvard Business Review, 29 January 2019 (https://hbr.org/2018/01/how-ai-could-help-the-public-sector, accessed 30 August 2020).

326. Digital technology, social protection and human rights: Report of the United Nations Special Rapporteur for extreme poverty. Geneva: Office of the High Commissioner for Human Rights; 2019 (https://www.ohchr.org/EN/Issues/Poverty/Pages/DigitalTechnology.aspx, accessed 21 March 2021).

327. Derrington D. Artificial intelligence for health and healthcare. McLean (VA): The MITRE Corporation; 2017 (https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf, accessed 17 August 2020).

328. Federal Trade Commission. Twitter; 2020 (https://twitter.com/FTC/status/1285578871803437057, accessed 15 March 2021).

329. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17 March 2018 (https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, accessed 23 November 2020).

330. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med. 2018;15[11]:e1002689.

331. AI Policy Observatory fact sheet. Paris: Organization for Economic Co-operation and Development; 2020 (https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf, accessed 17 November 2020).

332. Rao AS, Verweij G. Sizing the prize: What's the real value of AI for your business and can you capitalise? London: Pricewaterhouse Coopers; 2019 (https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf, accessed 24 August 2020).

333. Davis SLM. Perspective. The Trojan horse: Digital health, human rights, and global health governance. Health Human Rights J. 2020;22:41-8.

334. Artificial intelligence: Canada and France work with international community to support the responsible use of AI. Paris: Government of France; 2019 (https://www.gouvernement.fr/sites/default/files/locale/piece-jointe/2019/05/23_cedrico_press_release_ia_canada.pdf, accessed 13 September 2022).

335. The Global Partnership on Artificial Intelligence (https://gpai.ai/, accessed 13 September 2022).

336. Report of the Secretary-General: Roadmap for digital cooperation. New York City (NY): United Nations; 2020 (https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf, accessed 17 September 2020).

337. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. https://www.who.int/publications/i/item/9789240029200, accessed 13 September 2022).

Annex A

Considerations for the ethical design, deployment and use of artificial intelligence technologies for health

The following provides practical guidance for several key groups that use AI in the health field: AI designers and developers, ministries of health and health care institutions and providers. It reflects the main principles, ideas and recommendations in this document.

A.1 Considerations for AI developers

The following considerations are for individuals, research organizations and companies involved in the design, deployment and updating of AI technologies used in health. AI developers include professionals with expertise in computer science or AI, who often also have a background in clinical or health care. Some AI developers are not sited in health systems, even though the products they design will play an increasingly important role in health. Some providers and hospitals are investing in and designing AI technologies and should consider the issues listed below with their existing ethical obligations as medical providers.

Developers, research organizations and companies should consider systems to ensure that the values, principles and processes that guide their operations are aligned with the expectations of health systems.

The considerations listed below are not comprehensive but are steps that developers and companies should take to ensure that the technologies they design and deploy are used for the benefit of patients and providers. Three areas should be considered: the design, development and deployment of an AI technology, with further consideration of improving it after deployment.

A.1.1 Designing an AI technology

1 Clarify the objectives

An AI technology or tool can be used alone or as an integral part of a system. The intended uses, the values and the indirect outcomes for users should be clearly defined.

Specific considerations

– Define the intended uses and the expected outcomes.
– What are the main functions of the tool?
– Who will use the tool?
– How will it be used
– When and where will it be used or not used?
– Will there be secondary (indirect) users?
– How should the objectives and functions be prioritized according to the available resources?
– Will use of the tool have indirect outcomes?
– Are the validity and efficiency of the tool limited over time?

2 Engage multiple stakeholders and understand contexts

AI technologies used in health care depend on the context and must be designed to work appropriately for different types of health-care providers and different uses by patients or practitioners before, during or after clinical care.

Specific considerations

– Define all possible contexts in which the AI technology will be used, including geographical scope, users' background and main languages, digital skills and regulatory frameworks.
– Involve individuals who understand various contexts in design to align the objectives and expected outcomes and avoid transferring bias from the data and amplifying it.
– Design, discuss and validate the formulation, conceptualization, proposed approach and solution with stakeholders in the targeted settings, including policy- and decision-makers, project owners and leaders, project managers, solution engineers and developers, potential users, domain experts and experts in ethics and information privacy.
– Clearly delineate responsibilities during design, development and deployment and the conditions to be fulfilled for attribution of responsibility.
– Determine the operational and technical limitations to designing, developing, testing, using and maintaining the tool, including human resources, expertise and software and hardware requirements.

3 Define relevant ethical issues through consultation

Each AI technology will require consideration of ethical issues, such as bias, privacy, data collection and use and human autonomy (among the principles listed in section 5 of this document). Ethical concerns that often emerge during consultation should be identified and integrated into the design and development. (Recommendations for addressing bias and privacy, two ethical issues that are often relevant for the design of AI technologies for health, are discussed below.)

4 Assess risks

Risk assessment and mitigation are necessary in the design and development of technologies for use in human health. Risk should be assessed at each stage of development and reassessed regularly with stakeholders. The aim of developers should be for the AI technology to achieve the intended outcomes with a reduced level of risk. All major trade-offs should be clearly identified and considered.

Specific considerations

– What are the expected outcomes?
– What are the potential secondary and unexpected outcomes?
– What would be the impact and consequences of the unexpected outcomes?
– What are the available resources and potential trade-offs?
– What approaches would mitigate risk?

5 Address biases

Biases in data due to past or continuing discrimination could be replicated. An AI technology should be used only if such bias can be mitigated. AI should be designed to reduce inequities and bias.

Specific considerations

– Determine how the study data were collected and how new study data will be collected, and look for any bias in the data according to the context.
– Consider the majority and minority groups included in the data and whether any under-representation that results in bias can be mitigated.
– Examine the effects of ethnicity, age, race, gender and other traits, and ensure that AI technologies with biases do not have negative impacts on individuals and groups according to these different characteristics.
– Prepare effectively and demonstrably for post-implementation surveillance of the application.

6 Privacy by design and privacy by default

All possible steps should be taken to safeguard the security, privacy and confidentiality of the information used to develop and validate an AI technology in relevant contexts and of the information and data collected and produced by the AI technology.

Specific considerations

– Map the possible vulnerability of an AI technology with respect to privacy and reverse engineering in context.
– Identify data protection vulnerabilities in contracts and collaborations with (other) commercial parties and data-sharing systems and networks.
– Select design options that favour privacy and ensure that any reduction in privacy is consciously agreed to.
– Safeguard data protection and privacy preservation over time and with technology updates.

A.1.2 Developing an AI technology

1 Identify regulatory requirements

Regulatory frameworks for AI are evolving. While most regulatory frameworks address data protection, data security and privacy, emerging governance guidelines include equal access and human autonomy. Compliance measures should be included in development and updates of a technology.

Specific considerations

– Adhere to country-specific or regional export rules and guidelines, such as the EU GDPR, Singapore's Personal Data Protection Act or the US Health Insurance Portability and Accountability Act.
– Identify open concepts and open norms that should be specified for compliance, e.g., in GDPR Article 22, the "far reaching effects" in "Person may not be subjected solely to automated decision procedure with far reaching effects".
– Define relevant open norms and concepts that can be justified to affected parties and experts with relevant knowledge of the application.

2 Establish data management plans

Clear management plans and protection guidelines should be established for data collection, storage, organization and access to ensure data security and safeguard privacy and confidentiality.

Specific considerations

– Understand the data collection and sharing requirements and regulations in the countries, sectors and institutions of potential users, including legal requirements for managing consent for the use of training data.
– Determine the type of data that are being collected and where and how the data will be stored.
– Assess the physical infrastructure and operational processes that can be used to ensure data security and integrity.
– Understand and determine how confidentiality and privacy will be protected in different contexts.
– Establish guidelines and protocols for proper collection, storage, organization, access and use of personal, proprietary and public data in different contexts.
– Determine how long the data will be stored, when the data could be shared and other temporal considerations.

– Give preference to the use of anonymized data whenever possible.
– Determine who is responsible for data governance and ensure appropriate follow-up.
– Clearly identify all groups who will have access to the data throughout the product's life cycle.
– Determine any type of secondary use of data that could be allowed.

3 Adopt standards and best practices

Ensure the compliance and/or interoperability of the AI technology with other technologies that will be introduced into health systems. One or more established international, regional or national standards and/or performance benchmarks for an AI technology should be adopted according to regulations, guidance and application requirements, design and development plans.

Specific considerations (examples of standards)

– ISO standards (security and privacy)
– US National Institute of Standards and Technology (security and privacy)
– IEEE 7000 series (privacy and fairness)
– Health Level 7 (transfer of administrative and clinical health data)

A.1.3 Deploying an AI technology and improving it after deployment

1 Engage and educate multiple stakeholders for deployment and maintenance

Prioritize inclusivity throughout to ensure better understanding of needs and to build adapted solutions for multiple stakeholders.

Specific considerations

– Clearly delineate responsibility for what to do, when and how.
– Design, discuss and validate the proposed approach with various stakeholders in all targeted regions, including policy- and decision-makers, project owners and leaders, project managers, solution engineers and developers, potential users, domain experts and experts in ethics and information privacy.
– Train stakeholders in why, how and when to use the tool, including the main objectives, functions and features and differences among usage scenarios, when applicable.
– Engage continuously with stakeholders, and support users.

2 Evaluate and improve performance

The outcomes and impact on health care of the AI technology should be assessed formally, and the design and development of the technology continuously improved according to the ethical principles that initially guided its development and to new governance guidelines and all applicable legal obligations and regulations. The risks of the technology and of its intended usage in different health care settings should be assessed regularly to manage its deployment, continuous development and maintenance.

Specific considerations

The accuracy and risks of error of the AI technology should be evaluated to assess implications for:

– Incorporating, verifying and validating changes to the tool or system;
– monitoring and ensuring the effectiveness and usefulness of the tool or system over time;
– how long the results or the technology can be used;
– how often the tool or system should be updated; and
– who is responsible for updating.

A.2 Considerations for ministries of health

The following considerations are intended for ministries of health, which will have the primary responsibility for determining whether and how AI technologies should be integrated into health systems, the conditions under which they should be used, the protection of individuals that must accompany use of such technologies and policies that can address both expected and unexpected ethical challenges. Evaluation, regulation, deployment and oversight of AI technologies will require inter-ministerial coordination. Thus, while these considerations are directed to ministries of health, implementation will require collaboration with other relevant ministries, such as of information technology and education.

These considerations are not comprehensive but may be a starting-point for ministries of health to ensure that the use of AI technologies is consonant with the wider objective of the government to provide affordable, equitable, appropriate, effective health care, with the goal of attaining universal health coverage. Three areas should be considered: how ministries should protect the health and safety of patients, how they should prepare for the introduction and use of AI technologies and how they should address ethical and legal challenges and protect human rights.

A.2.1 How to protect the health and safety of patients

1 Assess whether AI technologies are appropriate and necessary

AI technologies should be used only if they are necessary and appropriate and contribute to achieving universal health coverage. They should not divert attention and resources from proven but underfunded interventions that would reduce morbidity and mortality.

Specific considerations

– Evaluate the institutional and regulatory context and infrastructure to determine whether the technology would be as cost-effective as "traditional" technologies and whether its introduction and use are in accordance with human rights.

– Conduct an impact assessment before deciding whether to implement or continue use of AI in the health system.

– Calculate the risk-benefit ratio of adoption, investment and uptake of an AI technology, and make the information available to stakeholders so that they can provide input to any evaluation or decision.

– Manage the ethical challenges of the AI technology (e.g., equitable access, privacy) appropriately.

2 Testing, monitoring and evaluation

AI must be rigorously tested, monitored and evaluated. Clinical trials can provide assurance that any unanticipated hazards or consequences of AI-based applications are identified and addressed (or avoided entirely) and an approved AI device can be re-tested and monitored to measure its performance and any changes that may occur once it has been approved.

Regulatory agencies can support testing, transparent communication of outcomes and monitoring of the performance and efficacy of a technology. Many LMIC still lack sufficient regulatory capacity to assess drugs, vaccines and devices, and the rapid arrival of AI technologies could mean that their regulatory agencies cannot accurately assess or regulate such technologies for the public good.

Specific considerations

– Countries should have sufficient regulatory capacity to ensure rigorous scrutiny of AI technologies on which countries rely in health care.

– For certain low-risk AI technologies, regulators may consider "lighter" premarket scrutiny.

– AI technologies should be tested prospectively in randomized trials and not against existing laboratory datasets.
– Regulatory scrutiny should be applied when data from non-health devices are imputed and used to train AI health technologies.

3 Assign liability

Reliance on AI technologies entails responsibility, accountability and liability and also compensation for any undue damage.

Specific considerations

– Ministry of health experts should evaluate AI tools to ensure accountability for any negative consequences that arise from their use.
– Liability rules used in clinical care and medicine should be modified to assess and assign liability, including product liability, the personal liability of decision-makers, input liability and liability to data donors. The rules should include causal responsibility, objective liability regimes and liability for retrospective harm as well as mechanisms for assigning vicarious liability when appropriate.

4 Ensure that all people are guaranteed redress in the legal system

Processes should be available for compensation of undue damage caused by use of AI technologies.

Specific considerations

– Independent oversight should be available to ensure equitable access to health care of appropriate quality.
– Swift, accessible mechanisms should be available for complaint, including for patients and health staff to demand protection of personal data and particularly of sensitive health data.

A.2.2 Prepare for the introduction and use of AI technologies

1 Institutional preparedness and technical capacity

Ministries of health should have the necessary human and technical resources to realize the full benefits of AI technologies for health while mitigating any negative impacts.

Specific considerations

– Training and capacity-building based on established criteria should be organized for government officials to evaluate whether an AI technology is based on ethical principles.
– Health-care authorities and medical professionals should be involved and engaged in AI design and, when possible, software engineering.
– Civil society, medical staff and patient groups should be consulted about the introduction of AI technology and included in both external audit and monitoring of its functioning.
– The introduction of an AI technology should be accompanied by appropriate investments by the health system to capture its benefits. For example, tools to predict a disease outbreak should be complemented by robust surveillance systems and other measures to respond effectively to an outbreak.

2 Infrastructure for AI technologies

The right infrastructure is a prerequisite for proper deployment of AI in a health-care system.

Specific considerations

– Criteria should be established to identify and measure the infrastructure requirements, including for operation, maintenance and oversight.

– When necessary, infrastructure should be provided or strengthened with civil society support and international cooperation.
– Ministries of health should identify effective alternatives if any infrastructure is lacking, if the AI technology is too expensive or if it poses a high risk to patients.

3 Management of data

Data must be of high quality to prevent unintended harm from use of AI systems, as limited, low-quality or inaccurate data could result in biased inferences, misleading data analyses and poorly designed applications for health. Other critical elements of health data management include protecting the privacy and confidentiality of patient data and the rules for sharing such data.

Specific considerations

– Data processing (including from non-medical devices) and its representativeness, accuracy, harmonization, accessibility, interoperability and reusability should be regulated, with the informed consent of data providers (patients).
– Access to and use of data from digital self-care applications and/or wearable technologies should also be regulated. Data from these applications and technologies should be collected, stored and used in accordance with principles for data minimization.
– Patients and consumers who provide data should have access to and be allowed to reuse and thereby benefit from their data. Their data should not be monopolized by an AI technology provider.
– Quality control measures should be implemented to ensure the representativeness of data from different population groups.
– Mechanisms and procedures should be in place to collect relevant patient data to train AI technology according to the environment, culture and specifics of the community in which the technology is intended to be used.
– Patients and consumers should know what data are used in training AI systems.

A.2.3 Address ethical and legal challenges and protect human rights

1 Preserve and enhance human autonomy

AI technologies for health should enhance human decision-making and empower medical professionals (clinicians and providers) rather than replace them.

Specific considerations

– Human judgement should be used with regard to prediction of disease and/or recommended treatment by an AI technology.
– Ministries of health should designate the types of information with which a clinician should be provided to make an independent judgement about an AI result or outcome.
– Meaningful, clear information should be provided to patients to allow them to make informed decisions about health recommendations based on AI technology.

2 Patient agency with regard to predictive algorithms

Use of AI predictive analytics in health care raises ethical concern with respect to informed consent and individual autonomy in decisions about patient and consumer health.

Specific considerations

– The need for an AI technology should be assessed, with the risk of the technology to patient autonomy and well-being.
– Patients should be allowed to refuse AI technologies for health.

– A mechanism should be available to inform patients of the benefits, risks, value, constraints, novelty and scope of an AI tool.

3 Privacy, confidentiality and informed consent in the collection and use of patient data

The autonomy and trust of patients who provide data are paramount, especially meaningful individual control over data. Health-data processing should include respect for the right to privacy and should ensure that patients maintain control over decisions, including their informed consent.

Specific considerations

– Up-to-date data protection and confidentiality laws should be a prerequisite for use of AI.
– Independent oversight and other forms of redress should be available to protect patient privacy and data confidentiality.
– Data protection supervisory agencies should have sufficient resources for effective privacy protection.
– Ministries of health should employ experts to determine whether AI tools meet standards of privacy to foster the general trust of patients who provide data.
– Ministries of health should have a protocol for collecting, storing and sharing personal data or data that could be identified and ensure that the data are managed in such a way as to protect privacy, including confidentiality and informed consent.
– Ministries of health should ensure that patients have the right to refuse data collection by and the data-sharing requirements of an AI technology. Explicit consent should be given for secondary uses of health data.
– Ministries of health should limit the collection of data to those required and not collect additional data.
– Ministries of health should provide training for health staff in the implications for the human rights of patients as part of capacity-building for use of AI technology.

4 Transparency of AI technologies for health

AI technologies must be provided and relied on transparently in order to assign responsibility and ensure trust and protection of patient rights.

Specific considerations

– Ministry of health experts should transparently evaluate an AI technology developed by others and make the results of such assessments publicly available throughout the life-cycle of the AI system.
– Ministries of health should ensure that clinicians can explain how an AI system has been validated to patients and their families.
– External experts should have enough information about the AI system and its training data to make independent assessments.

5 Ensure equitable access to AI technologies and related health care

When an AI technology is considered necessary (see above), ministries of health have an ethical obligation to ensure equitable access to that technology. Diagnostic use of AI should be extended carefully to avoid situations in which large numbers of people receive an accurate diagnosis of a health condition in the absence of appropriate treatment options.

Specific considerations

– Ministries of health have a duty to ensure equitable access to all to AI-based health care, regardless of gender, geography, ethnicity and other conditions.

– Ministries of health have a duty to provide treatment after AI-based testing and confirmation of disease.
– Ministries of health should ensure that the benefits of data from AI are fairly shared with the patients who provided the data for AI training and not monopolized by technology service providers.

A.3 Considerations for health-care institutions and providers

The following considerations are intended for health-care institutions and providers, such as hospitals, doctors and nurses. While programmers may be those primarily responsible for the design of AI technologies and ministries of health and regulatory agencies for approval and selection of such technologies for use, health-care providers determine which technologies to use and how and may also provide direct feedback to the health-care system, the medical community and the designers of the technologies about whether they meet the needs of patients.

The following is not comprehensive but may be used as a starting point as health-care providers increase use of AI for health care. Use of AI technologies for health outside regular health-care settings is discussed in section 3.1 of this document. Three areas are considered: whether the AI technology is necessary and appropriate; whether the context in which the AI technology will be used is appropriate; and whether a health-care provider should use a particular AI technology.

A.3.1 Is the AI technology necessary and appropriate?

1 Prioritize safety

Use of AI technology in health care will inadvertently address and could amplify risk-prone decisions, procedures or both. Technology-related risks must be counteracted by risk mitigation strategies, which should be integrated into AI decision-making or be applicable to AI decisions.

2 Promote transparency

Introduction of any AI technology must be sufficiently transparent that it can be criticized, by the public or by internal review mechanisms.

Specific considerations

– The source code should be fully disclosed.
– Algorithms must be open to criticism by an in-house or other appropriate expert.
– The data used to train the algorithm, whether certain groups were systematically excluded from such data, how the training data were labelled and by whom (including expertise and appropriateness of labelling) should be known.
– The underlying principles and value sets used for decision trees should be transparent.
– The learned code should be available for independent audit and review by appropriate third parties.

3 Address bias

Bias due to past or continuing discrimination could be replicated. An AI technology should be used only if such bias can be mitigated, and AI should be designed to reduce inequity and bias.

Specific considerations

– Ensure that AI with certain biases does not have negative impacts according to race or ethnicity or that the bias can be mitigated.
– If bias cannot be removed, ensure that this is stated transparently and reflected in decisions, e.g., to be taken into consideration by a provider or patient.

4 Safeguard privacy

Health-care providers must prevent re-identification, especially for datasets that can be linked by third parties to re-identify individuals.

Specific considerations

– Understand issues related to privacy and reverse engineering.

– Ensure that any option for use of an AI technology in a clinical setting favours privacy and that any reduction in privacy is actively agreed.

– Take the necessary measures to prevent leakage of identifiable information.

5 Institute regular challenge and review

Even if an AI technology is deemed appropriate up front, it must be subject to regular challenge and review. This may be necessary due to software erosion, changes in context over time and changes in the AI technology itself as it continues to learn from new data and evolves.

Specific considerations

– Establish regular technical review, including external review.

– Review whether the AI is having the intended impact, is filling a gap in need and is improving health care.

A.3.2 Is the context in which the AI technology will be used appropriate?

1 Assess whether the AI technology is necessary and appropriate in each clinical setting

Specific considerations

– Determine whether the AI technology offers advantages over what is currently offered and fills a gap.

– Compare the risks and benefits of the AI technology with those of current technology.

– Ensure that the AI technology is necessary and that the problem is clearly stated to ensure effective delivery of care that justifies use of the technology.

– Ensure that the AI technology is based on sufficient electronic health data.

– Ensure that the health data used were acquired in an ethical manner.

– Ensure the necessary infrastructure for use of the AI technology.

– Confirm the support of experts, including partnerships with academic institutions and commercial entities, and appropriate agreements with respect to IP, accountability, confidentiality, ethics, access and commercialization.

– Establish commonly agreed ethical principles for the collection, sharing and use of the data and its governance.

2 Understand local perspectives

The perspectives of local consumers should be recognized, particularly the sovereignty of indigenous peoples over their data for the collective benefit of people. This includes determining whether the health service has a "social license" to use AI, i.e., the consent of communities and/or individuals.

Specific considerations

– Public and consumer communication and education about AI should be adequate.

– Providers should secure a "social license" from the communities involved.

– Providers should ensure sovereignty and governance of indigenous populations over their data.

A.3.3 Should a health-care provider use the AI technology?

1 Ensure that the information provided by an AI technology can be interpreted

The information derived by an AI technology must be interpreted by a clinician. Human judgement is critical, and the context is important. Clinicians should be able understand the data and variables so that they can explain the principles of the AI application to themselves, colleagues, patients and families.

2 Understand the level of risk

Decisions made by clinicians on the basis of an AI technology must be transparent and based on understanding that they are appropriate or commensurate with any risk. AI should be used in prevention, treatment, rehabilitation and/or palliative care only if it the risk-benefit ratio is positive. It should not be used if the influence of the technology on risk is unclear or if it could increase or exacerbate risk. Specific guidelines for medical research involving human beings must be followed if AI technology is used experimentally.

3 Ensure responsible use of AI

Health-care providers must not only ensure that an AI technology is technically accurate but also consider whether it can be used responsibly. Health-care providers should state specifically why AI is appropriate in a particular situation.

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE DOCTOR-PATIENT RELATIONSHIP



Report commissioned by the
Steering Committee for Human Rights
in the fields of Biomedicine and Health (CDBIO)

Author: Brent Mittelstadt

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE DOCTOR-PATIENT RELATIONSHIP

By Brent Mittelstadt, Senior Research Fellow and Director of Research at the Oxford Internet Institute, University of Oxford, United Kingdom

All requests concerning the reproduction or translation of all or part of this document should be addressed to the Directorate of Communication (F-67075 Strasbourg Cedex).

All other correspondence concerning this document should be addressed to the Directorate General of Human Rights and Rule of Law.

© Council of Europe, December 2021

TABLE OF CONTENTS

1 ESSENTIAL ELEMENTS 4
2 INTRODUCTION 8
3 BACKGROUND AND CONTEXT 10
Common ethical challenges in AI 12
The Oviedo Convention and human rights principles regarding health 22
4 OVERVIEW OF AI TECHNOLOGIES IN MEDICINE 29
5 THEORETICAL FRAMEWORK OF THE DOCTOR-PATIENT RELATIONSHIP 35
Professional ethics in medicine 38
Fiduciary duties and the healing relationship 39
Emergent challenges in the doctor-patient relationship 41
6 POTENTIAL IMPACT OF AI ON THE DOCTOR-PATIENT RELATIONSHIP 44
Inequality in access to high quality healthcare 44
Transparency to health professionals and patients 45
Risk of social bias in AI systems 49
Dilution of the patient's account of well-being 51
Risk of automation bias, de-skilling, and displaced liability 52
Impact on the right to privacy 54
7 RECOMMENDATIONS FOR COMMON ETHICAL STANDARDS FOR TRUSTWORTHY AI 56
Intelligibility requirements for informed consent 57
Public register of medical AI systems for transparency 60
Collection of sensitive data for bias and fairness auditing 61
8 CONCLUDING REMARKS 64
Appendix: Medical virtues 66

1 ESSENTIAL ELEMENTS

1. In response to a call by the Committee on Bioethics (DH-BIO)[1] to work on trust, safety, and transparency, this report investigates the known and potential impacts of AI systems on the doctor-patient relationship. This impact is framed by the human rights principles referred to in the European Convention on Human Rights and Biomedicine of 1997, otherwise known as the "Oviedo Convention," and its subsequent amendments.

2. The deployment of AI in clinical care remains nascent. Clinical efficacy has been established for relatively few systems when compared to the significant research activity in healthcare applications of AI. Research, development, and pilot testing often do not translate into proven clinical efficacy, commercialization, or widespread deployment. The generalization of performance from trials to clinical practice generally remains unproven.

3. A defining characteristic of medicine is the 'healing relationship' between clinicians and patients. This relationship is augmented by the introduction of AI. However, the role of the patient, the factors that lead people to seek medical attention, and the patient's vulnerability are not changed by the introduction of AI as a mediator or provider of medical care. Rather, what changes is the means of care delivery, how it can be provided, and by whom. The shift of expertise and care responsibilities to AI systems can be disruptive in many ways.

4. The potential human rights impact of AI on the doctor-patient relationship can be categorised according to six themes: (1) Inequality in access to high quality healthcare; (2) Transparency to health professionals and patients; (3) Risk of social bias in AI systems; (4) Dilution of the patient's account of well-being; (5) Risk of automation bias, de-skilling, and displaced liability; and (6) Impact on the right to privacy.

5. Concerning (1), as an emerging technology the deployment of AI systems will not be immediate or universal across all member states or healthcare systems. Deployment across institutions and regions will inevitably be inconsistent in terms of scale, speed, and prioritisation.

6. The impact of AI on clinical care and the doctor-patient relationship remains uncertain and will certainly vary by application and use case. AI systems may prove to be more efficient than human care, but also provide lower quality care featuring fewer face-to-face interactions.

7. The inconsistent rollout of AI systems with uncertain impacts on access and care quality poses a risk of creating new health inequalities in member states.

[1] Committee replaced by the Steering Committee for Human rights in the fields of Biomedicine and Health (CDBIO).

8. Article 4 of the Oviedo Convention addresses care provided by healthcare professionals bound by professional standards. It remains unclear whether developers, manufacturers, and service providers for AI systems will be bound by the same professional standards.

9. Careful consideration must be given to the role played by healthcare professions bound by professional standards when incorporating AI systems that interact directly with patients.

10. Concerning (2), transparency and informed consent are key values in the AI-mediated doctor-patient relationship. The complexity of AI raises a question: how should AI systems explain themselves, or be explained, to doctors and patients? This question has many possible meanings: (i) How does an AI system or model function? How was a specific output produced by an AI system? (ii) How was an AI system designed and tested? How is it governed? (iii) What information is required to investigate the behaviour of AI systems? Answers to each of these questions may be necessary to achieve informed consent in AI-mediated care.

11. In cases where AI systems provide some form of clinical expertise, for example by recommending a particular diagnosis or interpreting scans, this requirement to explain one's decision-making would seemingly be transferred from doctor to AI system, or at least to manufacturer of AI system. The difficulty of explaining how AI systems turn inputs into outputs poses a fundamental challenge for informed consent. Aside from the patient's capacity to understand the functionality of AI systems, in many cases patients simply do not have sufficient levels awareness to make free and informed consent possible. AI systems use unprecedented volumes of data to make their decisions, and interpret these data using complex statistical techniques, both of which increase the difficulty and effort required to remain aware of the full scope of data processing and clinical analysis informing one's diagnosis and treatment.

12. AI systems interacting directly with patients should self-identify as an artificial system. Whether the usage of AI systems in care settings should always be disclosed to patients by clinicians and healthcare institutions is a more difficult question.

13. Concerning (3), AI systems are widely recognised as suffering from bias in their inputs, processing, and outputs. Biased and unfair decision-making often occurs not for technical or regulatory reasons, but rather reflects underlying social biases and inequalities. For example, samples in clinical trials and health studies have historically been biased towards white male subjects meaning results are less likely to apply to women and people of colour.

14. Social biases in AI systems can lead to unequal distribution of outcomes across patient populations and protected demographic groups. Western societies have long been marked by significant social inequality. These historical and contemporary trends influence the training of future systems. Without

intervention, these patterns in access to healthcare opportunities and resources will be learned and reinforced by AI systems.

15. Detecting biases in AI systems is not straightforward. Biased decision-making rules can be hidden in 'black box' models. Simply anonymising health data may not be an adequate solution to mitigate biases due to the influence of historical inequality and the existence of strong proxies for protected attributes (e.g., post code as a proxy for ethnicity). The various challenges of social bias, discrimination, and inequality suggest health professionals and institutions face a difficult task in ensuring their usage of AI systems does not further existing inequalities and create new forms of discrimination.

16. Concerning (4), the development of trust in a doctor-patient relationship may be inhibited by technological mediation. As a mediator placed between the doctor and patient, AI systems can inhibit tacit understanding of the patient's health and well-being and encourage both clinician and patient to discuss health solely in measurable quantities or machine interpretable terms.

17. Concerning (5), to ensure patient safety and replace the protection offered by human clinical expertise, robust testing and validation standards should be an essential pre-deployment requirement for AI systems in clinical care contexts. Evidence of clinical efficacy does not yet exist for many AI applications in healthcare, which has justifiably proven a barrier to widespread deployment.

18. Concerning (6), AI poses several unique challenges to the human right to privacy and complementary data protection regulations. These rights seek to provide individuals with greater transparency and control over automated forms of data processing. They will undoubtedly provide valuable protection for patients across a variety of use cases of medical AI.

19. The Oviedo Convention sets out a specific application of the right to privacy (Article 8 ECHR) which recognises the particularly sensitive nature of personal health information and sets out a duty of confidentiality for health care professionals.

20. Ethical standards need to be developed around transparency, bias, confidentiality, and clinical efficacy to protect patient interests in informed consent, equality, privacy, and safety. Such standards could serve as the basis for deployments of AI in healthcare that help rather than hinder the trusting relationship between doctors and patients.

21. Where AI can be observed to have a clear impact on rights and protections set out in the Oviedo Convention, it is appropriate for the Council of Europe to introduce binding recommendations and requirements for signatories concerning how AI is deployed and governed. Recommendations should focus on a higher positive standard of care with regards to the doctor-patient relationship to ensure it is not unduly disrupted by the introduction of AI in care settings.

22. The Council of Europe could set standards for what and how information about the recommendation of an AI system concerning a patient's diagnosis and treatment should be communicated to the patient. These standards should likewise address the doctor's role in explaining AI recommendations to patients and how AI systems can be designed to support the doctor in this role.

23. The capacity of AI to replace or augment human clinical expertise utilising highly complex analytics and unprecedented volumes and varieties of data suggests its impact on the doctor-patient relationship may be unprecedented.

24. The degree to which AI systems inhibit 'good' medical practice hinges upon the model of service. If AI is used solely to complement the expertise of health professionals bound by the fiduciary obligations of the doctor-patient relationship, the impact of AI on the trustworthiness and human quality of clinical encounters may prove to be minimal. At the same time, if AI is used to heavily augment or replace human clinical expertise, its impact on the caring relationship is more difficult to predict. It is entirely possible that new, broadly accepted norms for 'good' care will emerge through greater reliance on AI systems, with clinicians spending more time face-to-face with patients and relying heavily on automated recommendations. The impact of AI on the doctor-patient relationship nonetheless remains highly uncertain. We are unlikely to see a radical reconfiguration of care in the next five years in the sense of human expertise being replaced outright by artificial intelligence.

25. A radical reconfiguration of the doctor-patient relationship of the type imagined by some commentators, in which artificial systems diagnose and treat patients directly with minimal interference from human clinicians, continues to seem far in the distance.

26. Going forward, the ideal model of clinical care and AI deployment in healthcare is one that utilises the best aspects of human clinical expertise and AI diagnostics.

27. The doctor-patient relationship is a keystone of 'good' medical practice, and yet it is seemingly being transformed into a doctor-patient-AI relationship. The challenge facing AI providers, regulators, and policymakers is to set robust standards and requirements for this new type of 'healing relationship' to ensure patients' interests and the moral integrity of medicine as a profession are not fundamentally damaged by the introduction of AI.

2 INTRODUCTION

Technological solutions such as artificial intelligence (AI) are increasingly seen as a potential solution to growing resource pressures in medicine, healthcare, and biomedical research. AI systems promise novel means to evaluate and improve the quality of clinical care, undertake biomedical research and investigate new therapeutics and pharmaceuticals, and expand care offerings to previously underserved populations.[2] A key driver of innovation and adoption is the belief that AI may relieve health professionals from "certain time-consuming clerical tasks and could increase their time for caregiving practices."[3] Medical decision-making and care are increasingly supported by expert and robotics systems that assist in record management, diagnosis, treatment planning, and delivery of interventions. Home and social care are similarly transformed through the introduction of remote monitoring and management systems. Health can increasingly be monitored, modelled, and managed based on data representations of the patient, supplementing or replacing verbal accounts and face-to-face physical care.[4]

A unique impact of AI and other emerging data-intensive and algorithmic technologies is their capacity to augment and support human decision-making by recommending the best action to take in a given situation, the best interpretation of data, and so on.[5] But these systems can also be used to outright replace human decision-making, expertise, and face-to-face clinical care. Natural language processing applications such as OpenAI's GPT-3, for example, suggest a future in which initial patient contact and even triage can be handled in part or entirely by artificial conversational agents. AI systems are already used by clinicians and hospitals for clinical and operational decision-making, seen for instance in risk prediction, discharge planning, diagnostics, and decision-support systems.[6] Developments in deep learning likewise suggest a future in which drug discovery and biomedical research are increasingly driven by computational systems capable of intelligent behaviour.[7] Recent advances in the pharmaceuticals to treat a rare form of brain cancer or Deepmind's breakthrough in

[2] World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance* (2021); ITALIAN COMMITTEE FOR BIOETHICS, *Artificial Intelligence and Medicine: some ethical aspects* (2020), http://bioetica.governo.it/en/opinions/joint-opinions-icbicbbsl/artificial-intelligence-and-medicine-some-ethical-aspects/ (last visited Nov 30, 2021).
[3] COUNCIL OF EUROPE, *Artificial intelligence in health care: medical, legal and ethical challenges ahead* (2020).
[4] Brent Mittelstadt et al., *The Ethical Implications of Personal Health Monitoring*, 5 INTERNATIONAL JOURNAL OF TECHNOETHICS 37–60 (2014).
[5] George A. Diamond, Brad H. Pollock & Jeffrey W. Work, *Clinician decisions and computers*, 9 JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY 1385–1396 (1987); James G. Mazoué, *Diagnosis Without Doctors*, 15 J MED PHILOS 559–579 (1990).
[6] Rebecca Robbins & Erin Brodwin, *Patients aren't being told about the AI systems advising their care*, STAT (2020), https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/ (last visited Nov 9, 2021).
[7] World Health Organization, *supra* note 1.

protein folding via AlphaFold already show the potential of the state of the art in medical AI.[8]

While the promise of AI is clear, a significant area of uncertainty concerns its impact on the practice of healthcare, and in particular the doctor-patient relationship. Medical expertise is no longer the sole domain of trained medical professionals and researchers; rather, AI technologies create opportunities to provide care through a mix of public and private, professional and non-professional, and human and technological stakeholders.

In response to the growing recognition of these opportunities and risks of AI on the practice of medicine and clinical care by the Council of Europe, and the call by the Committee on Bioethics (DH-BIO) to work on trust, safety, and transparency in this context,[9] this report investigates the known and potential impacts of AI systems on the doctor-patient relationship. This impact is framed by the human rights principles referred to in the European Convention on Human Rights and Biomedicine of 1997 otherwise known as the "Oviedo Convention," and its subsequent amendments. Human rights principles regarding health may require certain standards to be met in the doctor-patient relationship which can be disrupted, displaced, or at least augmented by the usage of AI in clinical care.

The report is structured as follows:

- **Section 2** provides background and context concerning definitions of AI and related technologies, common ethical challenges posed by AI systems, and a brief historical overview of human rights principles regarding health in the context of the Oviedo Convention.
- **Section 3** reviews types of AI technologies in medicine, focusing in particular on AI systems aimed at augmenting clinical care and the patient experience.
- **Section 4** proposes a theoretical framework for the doctor-patient relationship based in human rights and connecting the aims of medicine to the standards of good medical practice as developed by medicine as a formal profession.
- **Section 5** then proposes several categories of current and potential impacts of AI systems on the doctor-patient relationship, focusing on issues of bias, inequality in access to care, opacity and transparency, patient autonomy and safety, clinician responsibility and automation bias, and the human right to privacy.
- **Section 6** concludes with recommendations aimed at bolstering human rights protections in the context of AI and the doctor-patient relationship.

[8] Diana M. Carvalho et al., *Repurposing vandetanib plus everolimus for the treatment of ACVR1-mutant diffuse intrinsic pontine glioma*, CANCER DISCOV (2021), https://cancerdiscovery.aacrjournals.org/content/early/2021/09/20/2159-8290.CD-20-1201 (last visited Nov 30, 2021); John Jumper et al., *Highly accurate protein structure prediction with AlphaFold*, 596 NATURE 583–589 (2021).
[9] COUNCIL OF EUROPE, *supra* note 2.

3 BACKGROUND AND CONTEXT

Concepts such as artificial intelligence (AI), machine learning, algorithm, and AI system have a wide array of meanings across academic, policy, and public discourse. Unhelpfully, the concepts are often used interchangeably.[10] For the sake of clarity, some definitions and distinctions will be offered.

Artificial intelligence refers to the demonstration of intelligence by a machine, wherein intelligence is understood in terms of its expression in humans and animals. As an academic field artificial intelligence studies "intelligent agents" or "computational intelligence", understood as systems that perceive their environment and take actions that maximize their chances of achieving their goals.[11] Machine learning can be understood as a specialised type of AI in which the agent, or computer program, improves its performance at some task through experience. Machine learning systems use "prior knowledge together with training data to guide learning."[12]

In simple terms, machine learning can be thought of as a type of software that learns from a training dataset, wherein labels are created and applied by human labellers according to prior knowledge. A classic example is an image recognition program which is taught to distinguish between classes of objects. In this case the training dataset would consist of a series of pre-labelled images from which the system can derive classification rules to apply to new images or datasets.

Algorithms can be understood as core components of machine learning and artificial intelligence systems that guide the processes of learning and turning input data into outputs. In mathematical terms an algorithm can be understood as a mathematical construct with "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions."[13] For clarity, a simpler definition can be offered: an algorithm is a well-defined sequence of steps that produce an output from some set of inputs.

A machine learning algorithm can be understood as a type of algorithm in which a part of the sequence of steps has been learnt rather than pre-defined. For example, a machine learning algorithm used for classification tasks develops classes that can generalise beyond the training data.[14] The algorithm creates a model to classify new inputs. A machine learning model is the internal data of the algorithm that is fitted to input data to improve performance.

Image recognition technologies, for example, can decide what types of objects appear in a picture. The algorithm 'learns' by defining rules to determine how new inputs will be classified. The model can be taught to the algorithm via hand labelled inputs (supervised learning); in other cases, the algorithm itself defines best-fit models to

[10] Robin K. Hill, *What an Algorithm Is*, 29 PHILOS. TECHNOL. 35–59, 36 (2015).
[11] David Poole, Alan Mackworth & Randy Goebel, *Computational Intelligence* (1998).
[12] Tom Mitchell, *Machine learning* (1997).
[13] Hill, *supra* note 9 at 47.
[14] Pedro Domingos, *A few useful things to know about machine learning*, 55 COMMUNICATIONS OF THE ACM 78–87 (2012).

make sense of a set of inputs (unsupervised learning).[15] In both cases, the algorithm defines decision-making rules to handle new inputs. Critically, a human user will typically not be able to understand the rationale of decision-making rules produced by the algorithm.[16]

Popular and policy definitions of these terms often do not follow these technical definitions which can cause confusion. The World Health Organization (WHO), for example defines artificial intelligence as "the performance by computer programs of tasks that are commonly associated with intelligent beings." Definitions of this type are on the one hand problematically broad, insofar as they turn on the definition of "intelligence" and scope of behaviours of "intelligent beings," and thus cannot be used to classify a particular system or AI or not-AI alone. With that said, the openness of the definition can also be helpful in policy terms by enabling additional systems to be captured beyond the state-of-the-art at the point of drafting.

Regardless of their limitations, policy definitions of AI are arguably more important than technical definitions if our concern is with harmonisation across regulatory and policy frameworks. The 'Artificial Intelligence Act' (AIA), a proposed horizontal risk-based regulatory framework proposed by the European Commission, offers a particularly broad definition of AI that promises to be an influential international policy going forward[17]:

> "'Artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Appendix I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."

Appendix of the AIA offers a non-comprehensive list of techniques and approaches that can be considered AI, which encompasses machine learning, logic and knowledge-based approaches, and a variety of statistical methods:

> "(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

[15] Bart W. Schermer, *The limits of privacy in automated profiling and data mining*, 27 COMPUTER LAW & SECURITY REVIEW 45–52 (2011); Martijn Van Otterlo, *A Machine learning view on profiling*, PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN–PHILOSOPHERS OF LAW MEET PHILOSOPHERS OF TECHNOLOGY 41–64 (2013).

[16] Andreas Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata*, 6 ETHICS INF TECHNOL 175–183, 179 (2004).

[17] EUROPEAN COMMISSION, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*, 2021/0106(COD) (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (last visited Oct 27, 2021).

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods."

As this definition shows the AIA's definition of 'AI system' does not align strictly with the technical definitions offered above. For example, in this definition machine learning is treated as a component of AI rather than as a specialised type of AI. To avoid ambiguity, we offer the following working definition of 'artificial intelligence system' for the purposes of this report:

> 'Artificial intelligence systems' refers to standalone or hardware-embedded software that acts as an intelligent agent or displays computational intelligence. An AI system can consist of one or more algorithms or models, but typically refers to complex systems in which multiple algorithms or models work together to perform a complex task.

Public discourse is currently dominated by concerns with a particular class of AI systems that make decisions and recommendations about important matters in life. These systems augment or replace analysis and decision-making by humans and are often used due to the scope or scale of data and rules involved. The number of features considered in classification tasks can run into the millions. This task replicates work previously undertaken by human workers, but on a much larger scale using qualitatively distinct decision-making logic. These systems make generally reliable (but not necessarily correct) decisions based upon complex rules that challenge or confound human capacities for action and comprehension.[18] In other words, this report addresses AI systems whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact.

Common ethical challenges in AI

Prior review of the ethical challenges facing AI has identified six types of concerns that can be traced to the operational parameters of decision-making algorithms and AI systems. The map reproduced and adapted in Figure 1 takes into account:

> "decision-making algorithms (1) turn data into evidence for a given outcome (henceforth conclusion), and that this outcome is then used to (2) trigger and motivate an action that (on its own, or when combined with other actions) may not be ethically neutral. This work is performed in ways that are complex and

[18] Brent Mittelstadt et al., *The ethics of algorithms: Mapping the debate*, 3 BIG DATA & SOCIETY (2016), http://bds.sagepub.com/lookup/doi/10.1177/2053951716679679 (last visited Dec 15, 2016). The remainder of Section 2.1 draws heavily from the findings and ethical framework proposed in this landscaping study.

(semi-)-autonomous, which (3) complicates apportionment of responsibility for effects of actions driven by algorithms."[19]

From these operational characteristics, three epistemological and two normative types of ethical concerns can be identified based on how algorithms process data to produce evidence and motivate actions. The proposed five types of concerns can cause failures involving multiple human, organisational, and technological agents. This mix of human and technological actors leads to difficult questions concerning how to assign responsibility and liability for the impact of AI behaviours. These difficulties are captured in traceability as a final, overarching, type of concern.

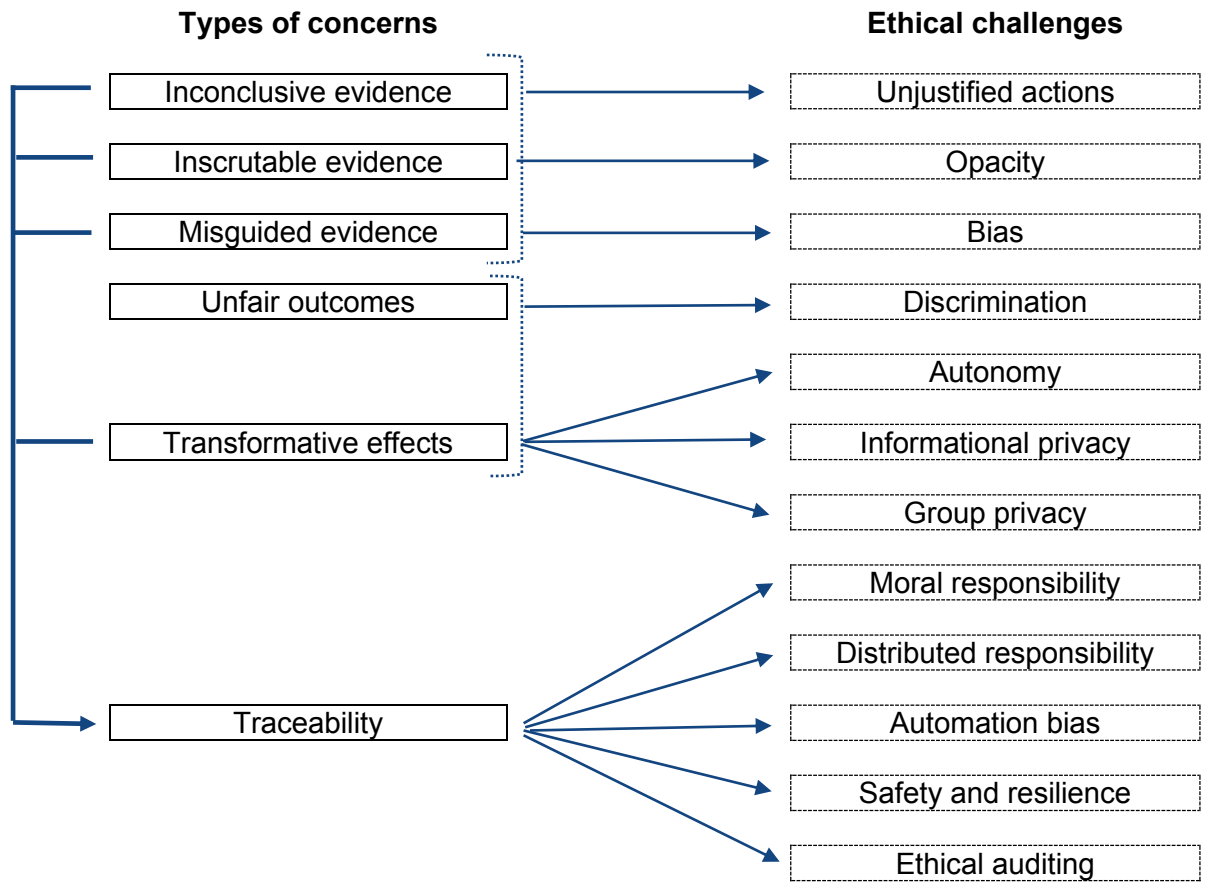


Figure 1 – Types of ethical concerns and challenges raised by algorithms (adapted from Mittelstadt et al., 2016)

The three aforementioned epistemological concerns with decision-making algorithms and AI systems can be defined as follows:

- **Inconclusive evidence** – When algorithms draw conclusions from the data they process using inferential statistics and/or machine learning techniques, they produce probable[20] yet inevitably uncertain knowledge. Statistical learning

[19] *Id.*

[20] The term 'probable knowledge' is used here in the sense of IAN HACKING, THE EMERGENCE OF PROBABILITY : A PHILOSOPHICAL STUDY OF EARLY IDEAS ABOUT PROBABILITY, INDUCTION AND STATISTICAL

theory[21] and computational learning theory[22] are both concerned with the characterisation and quantification of this uncertainty. Statistical methods can identify significant correlations, but correlations are typically not sufficient to demonstrate causality,[23] and thus may be insufficient to motivate action on the basis of knowledge of such a connection. The concept of an 'actionable insight' captures the uncertainty inherent in statistical correlations and normativity of choosing to act upon them.[24]

- **Inscrutable evidence** – When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be intelligible and open to scrutiny.[25] Given the complexity and scale of many AI systems, intelligibility and scrutiny cannot be taken for granted. A lack of access to datasets and the inherent difficulty of mapping how the multitude of data and features considered by an AI system contribute to specific conclusions and outputs cause practical as well as principled limitations.[26]

- **Misguided evidence** – Algorithms process data and are therefore subject to a limitation shared by all types of data processing, namely that the output can never exceed the input. The informal 'garbage in, garbage out' principle illustrates this phenomenon and its significance: conclusions can only be as reliable (but also as neutral) as the data they are based on.[27]

The three epistemic concerns detailed thus far address the quality of evidence produced by an algorithm that motivates a particular action. Normative concerns can be attached to these actions as well. There are two such potential normative concerns:

- **Unfair outcomes** – Algorithmically driven actions can be scrutinised from a variety of ethical perspectives, criteria, and principles. The normative acceptability of the action and its effects is observer-dependent and can be assessed independently of its epistemological quality. An action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence.

- **Transformative effects** – The impact of AI systems cannot always be attributed to epistemic or ethical failures. Much of their impact can appear initially ethically neutral in the absence of obvious harm. A separate set of

---

INFERENCE (2006). where it is associated with the emergence of probability and the rise of statistical thinking (for instance in the context of insurance) that started in the 17th Century.

[21] GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING (2013).

[22] Leslie G. Valiant, *A theory of the learnable*, 27 COMMUNICATIONS OF THE ACM 1134–1142 (1984).

[23] PETER GRINDROD, MATHEMATICAL UNDERPINNINGS OF ANALYTICS: THEORY AND APPLICATIONS (2014).

[24] Boaz Miller & Isaac Record, *Justified belief in a digital age: on the epistemic implications of secret Internet technologies*, 10 EPISTEME 117–134 (2013).

[25] Hilary Kornblith, *Epistemology: Internalism and Externalism* (2001).

[26] Miller and Record, *supra* note 23.

[27] For a formal approach to the 'garbage in, garbage out' principle, see: CLAUDE E. SHANNON & WARREN WEAVER, THE MATHEMATICAL THEORY OF COMMUNICATION (1998).

impacts, which can be referred to as transformative effects, concern subtle shifts in how the world is conceptualised and organised.[28]

A final overarching concern addresses the need to specify common characteristics of AI systems and environmental conditions to ensure accountability and liability can be fairly apportioned across all actors and stakeholders involved in developing, deploying, and using AI systems:

- **Traceability** – AI systems often involve multiple agents which can include human developers and users, manufacturers and deploying organisations, and the systems and models themselves. AI systems can also interact directly, forming multi-agent networks characterised by rapid behaviours that avoid the oversight and comprehension of their human counterparts due to speed, scale, and complexity. As suggested in the original landscaping study by Mittelstadt et al., "algorithms are software-artefacts used in data-processing, and as such inherit the ethical challenges associated with the design and availability of new technologies and those associated with the manipulation of large volumes of personal and other data."[29] All of these factors mean it is difficult to detect harms, find their cause, and assign blame when AI systems behave in unexpected ways. Challenges arising through any of the aforementioned five types of concerns can thus raise a related challenge concerning traceability, wherein both the cause and responsibility for bad behaviours need to be established.[30]

As detailed in Figure 1, these types of concerns with decision-making algorithms and AI systems can be traced to widely discussed ethical challenges and concepts. In brief, according to this approach the following are some of the key ethical challenges arising from operational characteristics of decision-making algorithms and the six types of concerns described above[31]:

- **Unjustified actions** – Much algorithmic decision-making and data mining relies on inductive knowledge and correlations identified within a dataset. Correlations based on a 'sufficient' volume of data are often seen as sufficiently credible to direct action without first establishing causality.[32] Acting on correlations can be

[28] LUCIANO FLORIDI, THE FOURTH REVOLUTION: HOW THE INFOSPHERE IS RESHAPING HUMAN REALITY (2014).
[29] Mittelstadt et al., *supra* note 17.
[30] G. O. Mohler et al., *Self-Exciting Point Process Modeling of Crime*, 106 JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 100–108 (2011); Luciano Floridi, *Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions*, 374 PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A: MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES 20160112 (2016).
[31] Note: this list is adapted from a literature review conducted by the author and reported in the following: Mittelstadt et al., *supra* note 17.
[32] Mireille Hildebrandt, *Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching*, 24 PHILOS. TECHNOL. 371–390 (2011); Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 THE MODERN LAW REVIEW 428–460 (2010); VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA : A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK AND THINK (2013); Tal Zarsky, *The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 118–132 (2016).

doubly problematic. Spurious correlations may be discovered rather than genuine causal knowledge. Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions with significant personal impact are directed towards individuals.[33]

- **Opacity** – This is the 'black box' problem with AI: the logic behind turning inputs into outputs may not be known to observers or affected parties or may be fundamentally inscrutable or unintelligible. Opacity in machine learning algorithms is a product of the high dimensionality of data, complex code and changeable decision-making logic.[34] Transparency and comprehensibility are generally desired because algorithms that are poorly predictable or interpretable are difficult to control, monitor and correct.[35] Transparency is often naively treated as a panacea for ethical issues arising from new technologies.[36]

Information about the functionality of algorithms is often intentionally poorly accessible.[37] Besides being accessible, information must be comprehensible to be considered transparent.[38] Efforts to make algorithms transparent face a significant challenge to render complex decision-making processes both accessible and comprehensible. The longstanding problem of interpretability in machine learning algorithms indicates the challenge of opacity in algorithms.[39] In the context of medicine, the World Health Organization (WHO) has recognized the critical importance of combatting opacity through provisions to ensure transparency, 'explainability', and intelligibility in the design and usage of AI in healthcare.[40]

- **Bias** – The automation of human decision-making is often justified by an alleged lack of bias in AI and algorithms.[41] This belief is unsustainable; AI

---

[33] PHYLLIS MCKAY ILLARI & FEDERICA RUSSO, CAUSALITY : PHILOSOPHICAL THEORY MEETS SCIENTIFIC PRACTICE (2014).
[34] Jenna Burrell, *How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOCIETY (2016).
[35] ANDREW TUTT, *An FDA for Algorithms* (2016), http://papers.ssrn.com/abstract=2747994 (last visited Apr 13, 2016).
[36] Anjanette Raymond, *The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics*, NORTHWESTERN JOURNAL OF INTERNATIONAL LAW & BUSINESS, FORTHCOMING (2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469291 (last visited Jul 22, 2015); Kate Crawford, *Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 77–92 (2016); Daniel Neyland, *Bearing Account-able Witness to the Ethical Algorithmic System*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 50–76 (2016).
[37] Tasha Glenn & Scott Monteith, *New Measures of Mental State and Behavior Based on Data Collected From Sensors, Smartphones, and the Internet*, 16 CURR PSYCHIATRY REP 1–10 (2014); Meredith Stark & Joseph J. Fins, *Engineering Medical Decisions*, 22 CAMBRIDGE QUARTERLY OF HEALTHCARE ETHICS 373–381 (2013); Rob Kitchin, *Thinking critically about and researching algorithms*, INFORMATION, COMMUNICATION & SOCIETY 1–16 (2016); Matthias Leese, *The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union*, 45 SECURITY DIALOGUE 494–511 (2014).
[38] Matteo Turilli & Luciano Floridi, *The ethics of information transparency*, 11 ETHICS INF TECHNOL 105–112 (2009).
[39] Hildebrandt, *supra* note 31; Leese, *supra* note 36; Burrell, *supra* note 33; TUTT, *supra* note 34.
[40] World Health Organization, *supra* note 1 at xiii.
[41] Engin Bozdag, *Bias in algorithmic filtering and personalization*, 15 ETHICS INF TECHNOL 209–227 (2013); Gauri Naik & Sanika S. Bhide, *Will the future of knowledge work automation transform personalized medicine?*, 3 APPLIED & TRANSLATIONAL GENOMICS 50–53 (2014).

systems unavoidably make biased decisions.[42] A system's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Development is not a neutral, linear path.[43] As a result, "the values of the author, wittingly or not, are frozen into the code, effectively institutionalising those values."[44] Inclusiveness and equity in both the design and usage of AI is thus key to combat implicit biases.[45] Friedman and Nissenbaum clarify that bias arise from (1) pre-existing social values found in the "social institutions, practices and attitudes" from which the technology emerges, (2) technical constraints and (3) emergent aspects of a context of use.[46]

- **Discrimination** – Discrimination against individuals and groups can arise from biases in AI systems. Discriminatory analytics can contribute to self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society.[47] While a single definition of discrimination does not exist, legal frameworks internationally have a long history of jurisprudence discussing types of discrimination (e.g., direct and indirect), goals of equality law (e.g., formal and substantive equality), and appropriate thresholds for distribution of outcomes across groups. In this context, embedding considerations of non-discrimination and fairness into AI systems is particularly difficult.[48] It may be possible to direct algorithms not to consider sensitive attributes that contribute to discrimination,[49] such as gender or ethnicity,[50] based

---

[42] Kevin Macnish, *Unblinking eyes: the ethics of automating surveillance*, 14 ETHICS INF TECHNOL 151–167 (2012); Sue Newell & Marco Marabelli, *Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'*, 24 THE JOURNAL OF STRATEGIC INFORMATION SYSTEMS 3–14, 6 (2015); Bozdag, *supra* note 40; Batya Friedman & Helen Nissenbaum, *Bias in computer systems*, 14 ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS) 330–347 (1996); Omer Tene & Jules Polonetsky, *Big data for all: Privacy and user control in the age of analytics* (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (last visited Oct 2, 2014); Felicitas Kraemer, Kees van Overveld & Martin Peterson, *Is there an ethics of algorithms?*, 13 ETHICS AND INFORMATION TECHNOLOGY 251–260 (2011).

[43] JEFFREY ALAN JOHNSON, *Technology and Pragmatism: From Value Neutrality to Value Criticality* (2006), http://papers.ssrn.com/abstract=2154654 (last visited Aug 24, 2015).

[44] Macnish, *supra* note 41 at 158.

[45] World Health Organization, *supra* note 1 at xiii.

[46] Friedman and Nissenbaum, *supra* note 41.

[47] Macnish, *supra* note 41; Leese, *supra* note 36; Solon Barocas, *Data Mining and the Discourse on Discrimination* (2014), https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf (last visited Dec 20, 2015).

[48] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*, 41 COMPUTER LAW & SECURITY REVIEW 105567 (2021); Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law*, 123 W. VA. L. REV. 735 (2020).

[49] SOLON BAROCAS & ANDREW D. SELBST, *Big Data's Disparate Impact* (2015), http://papers.ssrn.com/abstract=2477899 (last visited Oct 16, 2015).

[50] Toon Calders, Faisal Kamiran & Mykola Pechenizkiy, *Building classifiers with independency constraints*, *in* DATA MINING WORKSHOPS, 2009. ICDMW'09. IEEE INTERNATIONAL CONFERENCE ON 13–18 (2009); Faisal Kamiran & Toon Calders, *Classification with no discrimination by preferential sampling*, *in* PROC. 19TH MACHINE LEARNING CONF. BELGIUM AND THE NETHERLANDS (2010),

upon the emergence of discrimination in a particular context. However, proxies for protected attributes are not easy to predict or detect,[51] particularly when algorithms access linked datasets.[52]

- **Autonomy** – Value-laden decisions made by algorithms can also pose a threat to autonomy. Personalisation of content by AI systems, such as recommender systems, is particularly challenging in this regard. Personalisation can be understood as the construction of choice architectures which are not the same across a sample.[53] AI can nudge the behaviour of data subjects and human decision-makers by filtering information.[54] Different information, prices, and other content can be offered to profiling groups or audiences within a population defined by one or more attributes, for example the ability to pay, which can itself lead to discrimination. Personalisation reduces the diversity of information users encounter by excluding content deemed irrelevant or contradictory to the user's beliefs or desires.[55] This is problematic insofar as information diversity can be considered an enabling condition for autonomy.[56] The subject's autonomy in decision-making is disrespected when the desired choice reflects third-party interests above the individual's.[57]

A related challenge for autonomy concerns the intelligibility or comprehensibility of algorithmic systems and their outputs. Health professionals incorporating AI-based recommendations into their clinical care routines, for example, may experience a loss of autonomy if the basis for the recommendations is not well understood. Likewise, patients face a similar challenge when making informed decisions about their care based on AI recommendations. Recognising these risks, the WHO recognises "protecting human autonomy" as a key ethical principle for the design, usage, and governance of AI in healthcare due to the risk of decision-making power being transferred from humans to machines.[58]

- **Informational privacy and group privacy** – Algorithms also transform notions of privacy. Responses to discrimination, personalisation, and the inhibition of

---

http://wwwis.win.tue.nl/~tcalders/pubs/benelearn2010 (last visited Aug 24, 2015); Schermer, *supra* note 14.

[51] Zarsky, *supra* note 31; Andrea Romei & Salvatore Ruggieri, *A multidisciplinary survey on discrimination analysis*, 29 THE KNOWLEDGE ENGINEERING REVIEW 582–638 (2014).

[52] BAROCAS AND SELBST, *supra* note 48.

[53] Omer Tene & Jules Polonetsky, *Big data for all: Privacy and user control in the age of analytics*, NW.J. TECH. & INTELL. PROP. (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (last visited Oct 2, 2014).

[54] Mike Ananny, *Toward an Ethics of Algorithms Convening, Observation, Probability, and Timeliness*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 93–117 (2016).

[55] ELI PARISER, THE FILTER BUBBLE : WHAT THE INTERNET IS HIDING FROM YOU (2011); Belinda A. Barnet, *Idiomedia: The rise of personalized, aggregated content*, 23 CONTINUUM 93–99 (2009).

[56] Jeroen van den Hoven & Emma Rooksby, *Distributive justice and the value of information: A (broadly) Rawlsian approach*, 376 INFORMATION TECHNOLOGY AND MORAL PHILOSOPHY (2008).

[57] Stark and Fins, *supra* note 36; Sally A. Applin & Michael D. Fischer, *New technologies and mixed-use convergence: How humans and algorithms are adapting to each other*, *in* 2015 IEEE INTERNATIONAL SYMPOSIUM ON TECHNOLOGY AND SOCIETY (ISTAS) 1–6 (2015).

[58] World Health Organization, *supra* note 1 at xii.

autonomy due to opacity often appeal to informational privacy,[59] or the right of data subjects to "shield personal data from third parties." Informational privacy concerns the capacity of an individual to control information about herself,[60] and the effort required by third parties to obtain this information. A right to identity derived from informational privacy suggests that opaque or secretive profiling is problematic when carried out by a third party. In a healthcare setting this could include insurers, remote care providers (e.g., chatbot and triage service providers), consumer technology companies, and others. Opaque decision-making inhibits oversight and informed decision-making concerning data sharing.[61] Data subjects cannot define privacy norms to govern all types of data generically because the value or insightfulness of data is only established through processing.[62]

Privacy protections based upon identifiability are poorly suited to limit external management of identity via analytics. Current regulatory protections struggle to address the informational privacy risks of analytics owing to the definition of 'personal data' being linked to an identified or identifiable individual; identifying a user is often unnecessary for purposes of algorithmic profiling and decision-making. Rather, knowledge is generated about algorithmically curated groups rather than uniquely identifiable individuals. Existing regulatory frameworks for privacy and data protection do not reflect the importance of profiling and groups to modern data analytics and automated decision-making.[63]

- **Moral responsibility and distributed responsibility** – When a technology fails, blame and sanctions must be apportioned.[64] Blame can only be justifiably attributed when the actor has some degree of control and intentionality in carrying out the action.[65] Traditionally, developers and software engineers have had "control of the behaviour of the machine in every detail" insofar as they can explain its overall design and function to a third party.[66] This traditional conception of responsibility in software design assumes the developer can reflect on the technology's likely effects and potential for malfunctioning,[67] and

[59] Schermer, *supra* note 14.
[60] L. Van Wel & L. Royakkers, *Ethical issues in web data mining*, 6 ETHICS AND INFORMATION TECHNOLOGY 129–140 (2004).
[61] Hojung Kim, Joseph Giacomin & Robert Macredie, *A Qualitative Study of Stakeholders' Perspectives on the Social Network Service Environment*, 30 INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION 965–976 (2014).
[62] Van Wel and Royakkers, *supra* note 59; Hildebrandt, *supra* note 31.
[63] Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHILOSOPHY & TECHNOLOGY 475–494 (2017); 126 LINNET TAYLOR, LUCIANO FLORIDI & BART VAN DER SLOOT, GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES (2017), http://link.springer.com/book/10.1007/978-3-319-46608-8 (last visited Jan 18, 2017).
[64] Kraemer, van Overveld, and Peterson, *supra* note 41 at 251.
[65] Matthias, *supra* note 15.
[66] *Id.*
[67] Luciano Floridi, Nir Fresco & Giuseppe Primiero, *On malfunctioning software*, 192 SYNTHESE 1199–1220 (2014).

make design choices to choose the most desirable outcomes according to the functional specification.[68]

Justified allocation of moral responsibility is difficult for algorithms and AI systems with learning capacities. The traditional model for allocating responsibility in computing requires the system to be well-defined, comprehensible and predictable; complex and fluid systems (i.e., one with countless decision-making rules and lines of code) inhibit holistic oversight of decision-making pathways and dependencies. Machine learning algorithms are particularly challenging in this respect,[69] seen for instance in genetic algorithms that program themselves. The traditional model of responsibility fails because "nobody has enough control over the machine's actions to be able to assume the responsibility for them."[70] Distributed responsibility is thus a particular challenge for AI systems but could be addressed through application of strict liability or similar faultless responsibility schemes.

- **Automation bias** – A related problem concerns the diffusion of feelings of responsibility and accountability for users of AI systems, and the related tendency to trust the outputs of systems on the basis of their perceived objectivity, accuracy, or complexity.[71] Delegating decision-making to AI can shift responsibility away from human decision-makers. Similar effects can be observed in mixed networks of human and information systems as already studied in bureaucracies, characterised by reduced feelings of personal responsibility and the execution of otherwise unjustifiable actions.[72] Algorithms involving stakeholders from multiple disciplines can, for instance, lead to each party assuming others will shoulder ethical responsibility for the algorithm's actions.[73] Machine learning adds an additional layer of complexity between designers and actions driven by the algorithm, which may justifiably weaken blame placed upon the former.

- **Safety and resilience** – The need to apportion responsibility is acutely felt when algorithms malfunction. Unethical algorithms can be thought of as malfunctioning software artefacts that do not operate as intended. Useful distinctions exist between errors of design (types) and errors of operation (tokens), and between the failure to operate as intended (dysfunction) and the presence of unintended side-effects (misfunction).[74] Misfunctioning is distinguished from mere negative side effects by 'avoidability', or the extent to which comparable types of systems or artefacts accomplish the intended function without the effects in question. These distinctions clarify ethical aspects of AI systems that are strictly related to their functioning, either in the abstract

[68] Matthias, *supra* note 15.
[69] Burrell, *supra* note 33; Matthias, *supra* note 15; Zarsky, *supra* note 31.
[70] Matthias, *supra* note 15 at 177.
[71] Zarsky, *supra* note 31 at 121.
[72] HANNAH ARENDT, EICHMANN IN JERUSALEM: A REPORT ON THE BANALITY OF EVIL (1971).
[73] Michael Davis, Andrew Kumiega & Ben Van Vliet, *Ethics, Finance, and Automation: A Preliminary Survey of Problems in High Frequency Trading*, 19 SCIENCE AND ENGINEERING ETHICS 851–874 (2013).
[74] Floridi, Fresco, and Primiero, *supra* note 66.

(for instance when we look at raw performance), or as part of a larger decision-making system, and reveals the multifaceted interaction between intended and actual behaviour. Machine learning in particular raises unique challenges, because achieving the intended or "correct" behaviour does not imply the absence of errors or harmful actions and feedback loops.[75]

Both types of malfunctioning imply distinct responsibilities for algorithm and software developers, users and artefacts. Fair apportionment of responsibility for dysfunctioning and misfunctioning across large development teams and complex contexts of use is a difficult challenge. Requirements for resilience to malfunctioning as an ethical ideal in algorithm design need to be specified to ensure AI systems are both safe and resilient against dysfunctions and misfunctions. This reflects the ethical importance of human well-being and how it can be impacted by AI. Reflecting this, the WHO has explicitly recognized the importance of protecting human well-being and safety by enshrining it as a key ethical principle for usage of AI in healthcare.[76]

- **Ethical auditing** – How best to operationalise and set standards for testing of these ethical challenges remains an open question, particularly for machine learning. Merely rendering the code of an algorithm transparent is insufficient to ensure ethical behaviour. One possible path to achieve interpretability, fairness, and other ethical goals in AI systems is via auditing carried out by data processors,[77] external regulators,[78] or empirical researchers,[79] using ex post audit studies,[80] reflexive ethnographic studies in development and testing,[81] or reporting mechanisms designed into the algorithm itself.[82] For all types of AI, auditing is a necessary precondition to verify correct functioning. For systems with foreseeable human impact, auditing can create an ex post procedural record of complex automated decision-making to unpack problematic or inaccurate decisions, or to detect discrimination or similar harms.

[75] Except for trivial cases, the presence of false positives and false negatives in the work of algorithms, particularly machine learning, is unavoidable.
[76] World Health Organization, *supra* note 1 at xiii.
[77] Zarsky, *supra* note 31.
[78] TUTT, *supra* note 34; Zarsky, *supra* note 31; FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).
[79] Neyland, *supra* note 35; Kitchin, *supra* note 36.
[80] Christian Sandvig et al., *Auditing algorithms: Research methods for detecting discrimination on internet platforms*, DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY (2014), http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf (last visited Feb 13, 2016); Philip Adler et al., *Auditing Black-box Models by Obscuring Features*, ARXIV:1602.07043 [CS, STAT] (2016), http://arxiv.org/abs/1602.07043 (last visited Mar 5, 2016); Romei and Ruggieri, *supra* note 50; Kitchin, *supra* note 36; Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic investigation of computational power structures*, 3 DIGITAL JOURNALISM 398–415 (2015).
[81] Neyland, *supra* note 35.
[82] Alfredo Vellido, José David Martín-Guerrero & Paulo JG Lisboa, *Making machine learning models interpretable.*, 12 *in* ESANN 163–172 (2012).

The Oviedo Convention and human rights principles regarding health

The European Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine (ETS No. 164) of 1997, or the "Oviedo Convention," promotes the protection of human rights in biomedicine at a transnational level. The Oviedo Convention is a framework instrument meaning it contains general principles intended to be translated into domestic law by signatories. The Oviedo Convention contains many novel principles and requirements built on principles and rights contained in "previous international human rights treaties, such as the International Covenant on Civil and Political Rights of 1966 and the European Convention on Human Rights (ECHR) of 1950 (e.g. the rights to life, to physical integrity and to privacy, the prohibition of inhuman or degrading treatment and of any form of discrimination)."[83] The Oviedo Convention is inspired by and grounded in the rights to life, physical integrity and privacy, and prohibition of discrimination enacted through the ECHR. For the European Court of Human Rights, the Oviedo Convention has been used as an interpretative framework to elucidate and better understand the scope and significance of these rights in the context of biomedicine.[84]

The significance of these constituent human rights for the Oviedo Convention cannot be overstated. As a whole the Convention is designed to "protect the dignity and identity of all human beings and guarantee everyone, without discrimination, respect for their integrity and other rights and fundamental freedoms with regard to the application of biology and medicine" (Article 1). Across the Convention certain values and ends are explicitly upheld and protected, while others can be inferred through specific requirements. Above all, human dignity and the primacy of the patient are key to the Convention:

> "The notion of human dignity is clearly the bedrock of the Oviedo Convention. According to the Explanatory Report, "the concept of human dignity (...) constitutes the essential value to be upheld. It is at the basis of most of the values emphasised in the Convention." Recalling the history of the instrument, one of the members of the drafting group recognizes that "it was soon decided that the concept of dignity, identity and integrity of human beings/individuals should be both the basis and the umbrella for all other principles and notions that were to be included in the Convention.""[85]

Reference is made to other values and rights across the Oviedo Convention, such as the rights to life, physical integrity and privacy, and the prohibition of discrimination. For example, Article 10 reaffirms the right to privacy introduced in Article 8 of the

[83] Roberto Andorno, *The Oviedo Convention: A European Legal Framework at the Intersection of Human Rights and Health Law*, 2 133–143, 133 (2005).
[84] Francesco Seatzu & Simona Fanni, *The Experience of the European Court of Human Rights with the European Convention on Human Rights and Biomedicine*, 31 UTRECHT J. INT'L & EUR. L. 5–16 (2015).
[85] Andorno, *supra* note 82.

ECHR and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data:

> 1. "Everyone has the right to respect for private life in relation to information about his or her health.
>
> 2. Everyone is entitled to know any information collected about his or her health. However, the wishes of individuals not to be so informed shall be observed."

Following the transparency requirements implied by the right to privacy in Article 10, Article 5 of the Oviedo Convention affirms the well-established requirement for informed consent in medicine:

> "An intervention in the health field may only be carried out after the person concerned has given free and informed consent to it.
>
> This person shall beforehand be given appropriate information as to the purpose and nature of the intervention as well as on its consequences and risks.
>
> The person concerned may freely withdraw consent at any time."

According to the Explanatory Report, the requirement for consent "makes clear patients' autonomy in their relationship with health care professionals and restrains the paternalist approaches which might ignore the wish of the patient." Paragraphs 35 and 36 of the Report provide further details on the specific requirements for consent to be considered free and informed including constraints on the doctor's influence on a patient's decision and requirements concerning the quality, breadth, and clarity of information provided:

> "35. The patient's consent is considered to be free and informed if it is given on the basis of objective information from the responsible health care professional as to the nature and the potential consequences of the planned intervention or of its alternatives, in the absence of any pressure from anyone. Article 5, paragraph 2, mentions the most important aspects of the information which should precede the intervention but it is not an exhaustive list: informed consent may imply, according to the circumstances, additional elements. In order for their consent to be valid the persons in question must have been informed about the relevant facts regarding the intervention being contemplated. This information must include the purpose, nature and consequences of the intervention and the risks involved. Information on the risks involved in the intervention or in alternative courses of action must cover not only the risks

> inherent in the type of intervention contemplated, but also any risks related to the individual characteristics of each patient, such as age or the existence of other pathologies. Requests for additional information made by patients must be adequately answered.
>
> 36. Moreover, this information must be sufficiently clear and suitably worded for the person who is to undergo the intervention. The patient must be put in a position, through the use of terms he or she can understand, to weigh up the necessity or usefulness of the aim and methods of the intervention against its risks and the discomfort or pain it will cause."

Article 10 provides both a "right to know" and a "right not to know" about their health status and any information collected about their health. These rights are core elements of the doctor-patient relationship envisioned in the Oviedo Convention. If patients are entitled to make an informed decision about their care, it follows that they are entitled to receive adequate information to make that decision in an informed manner.[86]

Concerning discrimination, Article 11 explicitly prohibits discrimination on the grounds of genetic heritage. Likewise, Article 3 provides for equitable access to healthcare of an appropriate quality:

> "Parties, taking into account health needs and available resources, shall take appropriate measures with a view to providing, within their jurisdiction, equitable access to health care of appropriate quality."

Inequality in access to care or standards of care could be considered a violation of the prohibition on discrimination contained in Article 14 of the ECHR, in particular in relation to discrimination in "association with a national minority, property, birth or other status" (see section entitled "Inequality in access to high quality healthcare"). Similarly, Article 4 addresses quality of care and professional standards in healthcare and research:

> "Any intervention in the health field, including research, must be carried out in accordance with relevant professional obligations and standards."

The Oviedo Convention understandably does not specify quality standards to be met in healthcare and research, but rather leaves the determination of standards to professional bodies and domestic law of signatories of the Convention according to local health needs and available resources. With that said, as the Convention prescribes a minimum standard for human rights protections, member states can

[86] *Id.*

choose to enact higher standards in their translation of the Convention into domestic law. With regards to quality of care standards, this can be done in relation to Articles 3 and 4. Paragraph 30 of the Explanatory Report clarifies the parties envisioned as setting these professional obligations and standards:

> "30. All interventions must be performed in accordance with the law in general, as supplemented and developed by professional rules. In some countries these rules take the form of professional codes of ethics (drawn up by the State or by the profession), in others codes of medical conduct, health legislation, medical ethics or any other means of guaranteeing the rights and interests of the patient, and which may take account of any right of conscientious objection by health care professionals."

Paragraphs 31 and 32 elaborate on the nature of medicine as a profession, variation in standards across countries, the commitment of doctors to uphold ethical and legal standards, and the content and development of standards over time:

> "31. The content of professional standards, obligations and rules of conduct is not identical in all countries. The same medical duties may vary slightly from one society to another. However, the fundamental principles of the practice of medicine apply in all countries. Doctors and, in general, all professionals who participate in a medical act are subject to legal and ethical imperatives. They must act with care and competence, and pay careful attention to the needs of each patient.
>
> 32. It is the essential task of the doctor not only to heal patients but also to take the proper steps to promote health and relieve pain, taking into account the psychological well-being of the patient. Competence must be determined primarily in relation to the scientific knowledge and clinical experience appropriate to a profession or speciality at a given time. The current state of the art determines the professional standard and skill to be expected of health care professionals in the performance of their work. In following the progress of medicine, it changes with new developments and eliminates methods which do not reflect the state of the art. Nevertheless, it is accepted that professional standards do not necessarily prescribe one line of action as being the only one possible: recognised medical practice may, indeed, allow several possible forms of intervention, thus leaving some freedom of choice as to methods or techniques."

Following this, Paragraph 33 of the Explanatory Report provides a brief indication of the ideal model for the doctor-patient relationship with respect to choosing interventions:

> “33. Further, a particular course of action must be judged in the light of the specific health problem raised by a given patient. In particular, an intervention must meet criteria of relevance and proportionality between the aim pursued and the means employed. Another important factor in the success of medical treatment is the patient's confidence in his or her doctor. This confidence also determines the duties of the doctor towards the patient. An important element of these duties is the respect of the rights of the patient. The latter creates and increases mutual trust. The therapeutic alliance will be strengthened if the rights of the patient are fully respected.”

The Oviedo Convention thus specifies a number of rights and requirements relating to or derived from human rights protected in other contexts. Key values and interests can be derived from the topics addressed throughout the Convention. These values embedded in human rights principles regarding health can guide the development of a theoretical framework for the doctor-patient relationship. Specifically, the Oviedo Convention prescribes and discusses the following values:

- **Human dignity**
- **Primacy of patient interests over societal and scientific interests**
- **Right to life**
- **Physical integrity**
- **Privacy and identity**
- **Informed consent**
- **Right to know and right not to know**
- **Prohibition of discrimination and inequality in access to healthcare**
- **Quality of care standards**

In the section entitled “Theoretical framework of the doctor-patient relationship”, these values will be discussed in the context of the goals of medicine as a profession and societal good and used as the basis to develop a theoretical framework for the doctor-patient relationship. This framework, and the values underpinning it derived from the Convention, suggests that certain goods must be met in the doctor-patient relationship. Likewise, different models for clinical encounters and the doctor-patient relationship will align better or worse with these values. These issues will be picked up in the aforementioned section following a brief overview of AI systems in medicine.

To situate this report in ongoing policy work by the Council of Europe, it is important to briefly note recent reports that have addressed other areas of work relevant to the impact of AI in healthcare. The “Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 223)” was opened in October 2018 and is set to be ratified in October 2023. The Protocol amends Convention ETS No. 108. Of particular relevance to AI in medicine is its

revision of Article 8 (now Article 9) of the Convention to grant individuals a variety of data protection rights:

> 1. "Every individual shall have a right:
>
>   a. Not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration;
>   b. to obtain, on request, at reasonable intervals and without excessive delay or expense, confirmation of the processing of personal data relating to him or her, the communication in an intelligible form of the data processed, all available information on their origin, on the preservation period as well as any other information that the controller is required to provide in order to ensure the transparency of processing in accordance with Article 8, paragraph 1;
>   c. to obtain, on request, knowledge of the reasoning underlying data processing where the results of such processing are applied to him or her;
>   d. to object at any time, on grounds relating to his or her situation, to the processing of personal data concerning him or her unless the controller demonstrates legitimate grounds for the processing which override his or her interests or rights and fundamental freedoms;
>   e. to obtain, on request, free of charge and without excessive delay, rectification or erasure, as the case may be, of such data if these are being, or have been, processed contrary to the provisions of this Convention;
>   f. to have a remedy under Article 12 where his or her rights under this Convention have been violated;
>   g. to benefit, whatever his or her nationality or residence, from the assistance of a supervisory authority within the meaning of Article 15, in exercising his or her rights under this Convention."

Many of these rights mirror protections in the General Data Protection Regulation (GDPR), a data protection framework implemented by the European Commission in 2018, including a limited right not to be subject to an automated decision, a right to obtain information on data processing, and rights to request rectification and erasure of personal data.[87] These rights may come provide an important backbone to protect

[87] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INTERNATIONAL DATA PRIVACY LAW 76–99 (2017); Sandra Wachter & B. D. Mittelstadt, *A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI*, 2019 COLUMBIA BUSINESS LAW REVIEW (2019).

the ideal of informed consent in medical applications of AI by providing access to information about the scope and nature of automated processing.

The October 2020 report "Artificial intelligence in health care: medical, legal and ethical challenges ahead," published by the Parliamentary Assembly of the Council of Europe and drafted by its Committee on Social Affairs, Health and Sustainable Development, proposed a draft recommendation responding to the growing impact of AI in healthcare.[88] The report's explanatory memorandum discusses in great detail the various medical, legal, and ethical impacts envisioned for AI, which include:

- **Need for ethical review in biomedical research and limitations on competences and capacities of ethics review bodies to assess the unique risks and opportunities of AI**
- **Liability of AI providers in medicine and healthcare**
- **Protection of personal data in the context of harmonising data systems and supporting AI innovation and research in Europe, in particular**
- **Ensuring lawfulness, fairness, purpose specification, proportionality, privacy-by-design and default, responsibility, compliance, transparency, data security, and risk management**
- **Challenges of guaranteeing meaningful control and informed consent for patients and other data subjects**
- **Positive obligations for states to protect life and health via national reporting mechanisms**
- **Navigating the tension between "freedom to innovate" and meaningful protection of human rights**

Rather than being discussed in detail here, these and other points raised in prior reports from the Council of Europe are reflected in the discussion of potential impacts on the doctor-patient relationship in the section entitled "Potential impact of AI on the doctor-patient relationship".

[88] COUNCIL OF EUROPE, *supra* note 2.

4 OVERVIEW OF AI TECHNOLOGIES IN MEDICINE

As described in the section entitled "Background and context", a broad array of technologies can be described as AI. With high-level definitions of relevant concepts including artificial intelligence, algorithms, and machine learning are defined, it is necessary to explore in more detail the potential types of medical AI applications. As this report focuses on the impact of AI on the doctor-patient relationship, not all potential medical applications will be considered. As a first step, we can distinguish between three types of AI according to their intended users:

- **AI for biomedical researchers**
- **AI for patients**
- **AI for health professionals**

> Of these categories, AI for patients and health professionals are most relevant for the purposes of this report given the focus on the doctor-patient relationship.

Other taxonomies are of course possible; a recent report by the WHO, for example, distinguishes between AI applications for use in:

- **Health care**
- **Health research and drug development**
- **Health systems management and planning**
- **Public health and public health surveillance**

The taxonomy deployed here focuses on the intended users of AI systems because appropriate solutions to ethical challenges introduced by these systems typically vary according to the interests, level of expertise, and requirements of different stakeholder groups.

Although not directly relevant to the doctor-patient relationship, it is worth reviewing a few examples of AI used for medical research. One of the most common applications in biomedical research is drug discovery. For example, a recent discovery by computer scientists and cancer specialists at the Institute of Cancer Research and Royal Marsden NHS Foundation Trust of a new drug regime for a rare form of brain cancer in children (diffuse intrinsic pontine glioma).[89] Deepmind's recent advances on protein folding via AlphaFold likewise indicate the promise of AI for fundamental research.[90]

[89] Andrew Gregory, *Scientists use AI to create drug regime for rare form of brain cancer in children*, THE GUARDIAN, September 22, 2021, https://www.theguardian.com/science/2021/sep/23/scientists-use-ai-to-create-drug-regime-for-rare-form-of-brain-cancer-in-children (last visited Sep 26, 2021); Carvalho et al., *supra* note 7.
[90] Jumper et al., *supra* note 7.

AI can also be used for structuring, labelling, and searching unorganized or heterogenous medical datasets; image classifiers, for example, can process huge volumes of medical imaging data much faster than manual labellers. Such systems can also be useful for administrative and operational purposes as discussed below.

One noteworthy usage of AI that blurs the boundaries between research and clinical care is that of polygenic embryo screening, in which an algorithm summarizes "the estimated effect of hundreds or thousands of genetic variants associated with an individual's risk of having a particular condition or trait." This practice raises the spectre of eugenics by potentially allowing parents to select embryos both for health advantages, but also for socially desirable non-disease-related traits.[91]

Many AI applications are in development to be used directly by patients, often in collaboration with a health professional or artificial agent. These include telemedicine applications used for remote observation, clinical encounters, and video-observed therapy; virtual assistants and chat bots for information or triage; applications for managing chronic illnesses such as cardiovascular disease or hypertension; health and well-being 'apps'; personal health monitoring systems including wearables with built-in analytics and behavioural recommendations; and remote monitoring systems for facial recognition, gait detection, biometrics, and health-related behaviours.[92]

One purported benefit of AI systems aimed at patients it to "empower patients and communities to assume control of their own health care and better understand their evolving needs."[93] Health monitoring and telemedicine systems could, for example, assist patients in self-management of chronic conditions like diabetes, hypertension, or cardiovascular disease.[94] Therapeutic "chat bots" may also be able to assist in management of mental health conditions.[95] It has been predicted, for example, that the GPT-3 natural language application could eventually be used as the basis for conversational agents working directly with patients, for example as an initial point of contact or (more controversially) for triaging non-critical patients.[96] These applications seem highly likely given the existing deployment of 'virtual GP' chat bots which direct service enquiries and provide information to patients[97]; it should be noted, however, that such applications have been the subject of significant debate over their ethical

[91] Sheetal Soni & Julian Savulescu, *Polygenic Embryo Screening: Ethical and Legal Considerations*, THE HASTINGS CENTER (2021), https://www.thehastingscenter.org/polygenic-embryo-screening-ethical-and-legal-considerations/ (last visited Nov 23, 2021).
[92] Mittelstadt et al., *supra* note 3.
[93] World Health Organization, *supra* note 1.
[94] Mittelstadt et al., *supra* note 3; SECRETARY OF STATE FOR HEALTH AND SOCIAL CARE, *The Topol Review: Preparing the healthcare workforce to deliver the digital future* (2019), https://topol.hee.nhs.uk/.
[95] SECRETARY OF STATE FOR HEALTH AND SOCIAL CARE, *supra* note 93.
[96] Diane M. Korngiebel & Sean D. Mooney, *Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery*, 4 NPJ DIGITAL MEDICINE 1–3 (2021).
[97] Weiyu Wang & Keng Siau, *Trust in health chatbots* (2018); Claire Woodcock et al., *The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study*, 23 JOURNAL OF MEDICAL INTERNET RESEARCH e29386 (2021).

acceptability and regulation.[98] Likewise, they may lead to reduced access to human care.[99]

Finally, a wide variety of applications are aimed at health professionals. Three broad categories can be distinguished:

- **Applications designed for diagnostics, therapeutics, and other forms of clinical care**
- **Applications designed for operational or administrative uses**
- **Applications designed for public health surveillance**

The distinction between these categories is not always clear, as will be discussed below. To limit the focus of this report to the potential impact of AI on the doctor-patient relationship, only the first two categories will be surveyed. Public health surveillance could also be conceived as an extension of the clinical experience or doctor-patient relationship, insofar as patients may be contacted proactively by public health officials for clinical follow-up. Nonetheless, this report is concerned principally with the immediate clinical experience and relationship between individual health professionals and their patients.

AI systems aimed at clinical care are designed to fulfil a broad range of tasks, including diagnosis recommendations, optimization of treatment plans, and various other forms of decision-support.

According to the WHO:

> "AI is being evaluated for use in radiological diagnosis in oncology (thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging and dose optimization for radiological treatment), in non-radiological applications (dermatology, pathology), in diagnosis of diabetic retinopathy, in ophthalmology and for RNA and DNA sequencing to guide immunotherapy."[100]

Future applications currently in development (but not yet deployed clinically) include systems to detect "stroke, pneumonia, breast cancer by imaging,[101] coronary heart

[98] GARETH IACOBUCCI, ROW OVER BABYLON'S CHATBOT SHOWS LACK OF REGULATION (2020); Wang and Siau, *supra* note 96.
[99] World Health Organization, *supra* note 1.
[100] Wenya Linda Bi et al., *Artificial intelligence in cancer imaging: clinical challenges and applications*, 69 CA: A CANCER JOURNAL FOR CLINICIANS 127–157 (2019); World Health Organization, *supra* note 1.
[101] Pranav Rajpurkar et al., *Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists*, 15 PLOS MEDICINE e1002686 (2018); Babak Ehteshami Bejnordi et al., *Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer*, 318 JAMA 2199–2210 (2017).

disease by echocardiography[102] and detection of cervical cancer,"[103] including systems designed specifically for use in low- and middle-income countries (LMIC).[104] Systems are being designed to predict the risk of lifestyle diseases including cardiovascular disease[105] and diabetes.[106]

Development of medical image classification systems has been highly prevalent in recent years. Prior work, for example, has shown that neural networks can achieve consistently higher sensitivity for pathological findings in radiology.[107] Image classification systems can also be used to support detection of tuberculosis,[108] COVID-19, and other conditions through interpreting staining images[109] or X-rays.[110] Another emerging phenomenon is that of "digital twins," which are systems that simulate individual organs or multi-organ systems of individual patients for purposes of disease modelling and prediction.[111]

Generally speaking, the deployment of AI in clinical care remains nascent. Clinical efficacy has been established for relatively few systems when compared to the significant research activity in healthcare applications of AI. Research, development, and pilot testing often do not translate into proven clinical efficacy, commercialization, or widespread deployment. The generalization of performance from trials to clinical practice generally remains unproven.[112]

A 2019 meta-analysis of deep-learning image classifiers in healthcare found that despite claims of equivalent accuracy between AI systems and human healthcare professionals:

---

[102] Maryam Alsharqi et al., *Artificial intelligence and echocardiography*, 5 ECHO RESEARCH AND PRACTICE R115–R125 (2018).
[103] Using Artificial Intelligence to Detect Cervical Cancer, , NIH DIRECTOR'S BLOG (2019), https://directorsblog.nih.gov/2019/01/17/using-artificial-intelligence-to-detect-cervical-cancer/ (last visited Dec 1, 2021).
[104] World Health Organization, *supra* note 1; Innovative, affordable screening and treatment to prevent cervical cancer, , UNITAID , https://unitaid.org/project/innovative-affordable-screening-and-treatment-to-prevent-cervical-cancer/ (last visited Dec 1, 2021).
[105] Rui Fan et al., *AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus*, 10 SCIENTIFIC REPORTS 1–8 (2020); Yang Yan et al., *The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine?*, 16 JOURNAL OF GERIATRIC CARDIOLOGY: JGC 585 (2019).
[106] Jyotismita Chaki et al., *Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review*, JOURNAL OF KING SAUD UNIVERSITY-COMPUTER AND INFORMATION SCIENCES (2020).
[107] Ohad Oren, Bernard J Gersh & Deepak L Bhatt, *Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints*, 2 THE LANCET DIGITAL HEALTH e486–e488 (2020).
[108] Yan Xiong et al., *Automatic detection of mycobacterium tuberculosis using artificial intelligence*, 10 JOURNAL OF THORACIC DISEASE 1936 (2018).
[109] *Id.*
[110] Apoorva Mandavilli, *These Algorithms Could Bring an End to the World's Deadliest Killer*, THE NEW YORK TIMES, November 20, 2020, https://www.nytimes.com/2020/11/20/health/tuberculosis-ai-apps.html (last visited Dec 1, 2021).
[111] Matthias Braun, *Represent me: please! Towards an ethics of digital twins in medicine*, J MED ETHICS (2021).
[112] World Health Organization, *supra* note 1 at 6.

> "Few studies present externally validated results or compare the performance of deep learning models and health-care professionals using the same sample." Likewise, "poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy."[113]

The evidence base for clinical efficacy of deep learning systems may have improved in subsequent years, but broad adoption will seemingly hinge on standardised reporting of accuracy to enable assessment of clinical efficacy by medical regulators and clinical care excellence bodies.

A near term challenge for image classifiers is to build systems which can assess multiple image or scan types, such as X-rays and CT scans, which are often considered in combination by human radiologists while AI systems typically can only interpret one or the other. A similar challenge exists for detection of multiple conditions or pathologies, with existing classifiers often trained to only detect a single type of abnormality.[114]

Finally, many AI systems are also designed for administrative or operational purposes. AI systems can help with several aspects of hospital administration and operational evaluations. Discharge planning tools, for instance, can estimate discharge dates and barriers for hospitalized patients and flag up those that are clinically (nearly) ready to be discharged to clinicians, along with a list of necessary steps to take prior to discharge. Some systems can even schedule necessary follow-up appointments and care.[115] Natural language processing systems could be used for automation of routine or labour-intensive tasks, such as searching and navigation of electronic health record (EHR) systems or automated preparation of medical documentation and orders.[116] According to the WHO, "Clinicians might use AI to integrate patient records during consultations, identify patients at risk and vulnerable groups, as an aid in difficult treatment decisions and to catch clinical errors. In LMIC, for example, AI could be used in the management of antiretroviral therapy by predicting resistance to HIV drugs and disease progression, to help physicians optimize therapy."[117]

Distinguishing between uses of AI for clinical care and research versus those used for operational and quality improvement purposes by hospitals and health systems is often difficult. Many such systems are designed to identify at-risk patients. The UCLA Health network, for example, uses a tool that identified patients in primary care that are at high risk of being hospitalized or making frequent visits to an emergency room in the coming year. Similarly, Oregon Health and Science University use a regression

---

[113] Xiaoxuan Liu et al., *A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis*, 1 THE LANCET DIGITAL HEALTH e271–e297 (2019).
[114] Stephanie Price, *Technological innovations of AI in medical diagnostics*, HEALTH EUROPA (2020), https://www.healtheuropa.eu/technological-innovations-of-ai-in-medical-diagnostics/103457/ (last visited Sep 6, 2021).
[115] Robbins and Brodwin, *supra* note 5.
[116] Korngiebel and Mooney, *supra* note 95.
[117] World Health Organization, *supra* note 1; Jerome Amir Singh, *Artificial Intelligence and global health: opportunities and challenges*, 3 EMERGING TOPICS IN LIFE SCIENCES 741–746 (2019).

algorithm to monitor patients across the hospital for signs of sepsis.[118] Both are treated as a type of operational tool for monitoring and prioritising quality of care, and not as part of clinical care or research.

[118] Robbins and Brodwin, *supra* note 5.

5 THEORETICAL FRAMEWORK OF THE DOCTOR-PATIENT RELATIONSHIP

Health is a fundamental good valued across many contexts, including personal, social and economic life, related to the maintenance and well-being of the whole person. Without health personal plans cannot be made, projects pursued, or identities created without restrictions imposed by a physical, mental or social ailment.[119] Health is therefore a prerequisite for the realisation of other human goods.

Broadly speaking, the end of medicine is to guarantee the health of a society and individuals within it.[120] Despite the difficulties of defining health and illness as concepts, medicine is broadly recognised as a practice to promote health, thereby working towards a fundamental good.[121] A lack of agreement on a 'correct' definition of health, reflected in debate on the topic, does not undermine the fundamental value of health to human life.[122] The ends of medicine are achieved through 'good' medical encounters with individual patients.[123] In pursuing these ends in the doctor-patient relationship, moral and technical capacities must work together in the interests of the patient because medical activity affects individuals with moral worth and interests.

As discussed in the section entitled "The Oviedo Convention and human rights principles regarding health", the Oviedo Convention prescribes the following values:

- **Human dignity**
- **Primacy of patient interests over societal and scientific interests**
- **Right to life**
- **Physical integrity**
- **Privacy and identity**
- **Informed consent**

[119] Andrew Edgar, *The expert patient: Illness as practice*, 8 MEDICINE, HEALTH CARE AND PHILOSOPHY 165–171 (2005).
[120] WORLD HEALTH ORGANIZATION, *Preamble to the Constitution of the World Health Organization* (1948); KENNETH WILLIAM MUSGRAVE FULFORD, MORAL THEORY AND MEDICAL PRACTICE (1989).
[121] FULFORD, *supra* note 119; EDMUND D PELLEGRINO & DAVID C THOMASMA, THE VIRTUES IN MEDICAL PRACTICE (1993); Paul Schotsmans, Bernadette Dierckx de Casterle & Chris Gastmans, *Nursing considered as moral practice: a philosophical-ethical interpretation of nursing*, 8 KENNEDY INSTITUTE OF ETHICS JOURNAL 43–69 (1998).
[122] FULFORD, *supra* note 119; Alan Petersen, *Risk, governance and the new public health*, *in* FOUCAULT: HEALTH AND MEDICINE 189–206 (Alan Petersen & Robin Bunton eds., 1997); Adele E. Clarke et al., *Biomedicalization: Technoscientific transformations of health, illness, and U.S. biomedicine*, 68 AMERICAN SOCIOLOGICAL REVIEW 161–194 (2003).
[123] ALASDAIR MACINTYRE, AFTER VIRTUE: A STUDY IN MORAL THEORY (3rd Revised edition ed. 2007); PELLEGRINO AND THOMASMA, *supra* note 120; GENERAL MEDICAL COUNCIL, *Good Medical Practice* (2013), http://www.gmc-uk.org/static/documents/content/GMP_2013.pdf_51447599.pdf.

- **Right to know and right not to know**
- **Prohibition of discrimination and inequality in access to healthcare**
- **Quality of care standards**

These values, and the different goals of medicine as a practice, can be realised through different types of doctor-patient relationships. Models of the (ideal) doctor-patient relationship have adapted over time in recognition of the growing importance of patient autonomy and its appropriate balance with other ethical obligations of the doctor towards beneficence, non-maleficence, and justice.[124] An influential paper from Emanuel and Emanuel (1992) proposed four models for the doctor-patient relationship:

- **Paternalistic Model** – This model vests the vast majority of decision-making power in the doctor. It assumes the existence of shared, objective values or criteria to define the best course of action to promote the patient's health and well-being. The doctor's role is expert, skilled practitioner tasked with "promoting the patient's well-being independent of the patient's current preferences." The doctor acts as "the patient's guardian, articulating and implementing what is best for the patient." Autonomy is realised only through patient assent to the doctor's determination of the best course of action.
- **Informative Model** – In contrast, this model vests the vast majority of decision-making power in the patient. The objective of clinical interactions "is for the doctor to provide the patient with all relevant information, for the patient to select the medical interventions he or she wants, and for the doctor to executive the selected interventions." Objectives values are not assumed; rather, the patient's values and interests are taken as known or fixed to the patient but not the doctor. The doctor's role is to provide facts to facilitate the patient making a decision that bests matches their interests.
- **Interpretive Model** – This model closely follows the informative model but provides a greater role for the doctor to assist the patient in understanding her values and interests, and the possible impact of different interventions in these terms. The doctor acts as an advisor to help the patient "elucidate and make coherent" their values but does not pass judgement on these values or attempt to prioritize them on behalf of the patient. The ultimate choice of intervention still rests with the patient in the interpretive model, but the doctor plays a more active role in shaping this choice than the informative model.
- **Deliberative Model** – This model closely follows the interpretive model but gives the doctor a greater role in judging and prioritizing patient values. It is the doctor's role to "elucidate the types of values embodied in the available

[124] TOM L. BEAUCHAMP & JAMES F. CHILDRESS, PRINCIPLES OF BIOMEDICAL ETHICS (2009); E. J. Emanuel & L. L. Emanuel, *Four models of the physician-patient relationship*, 267 JAMA: THE JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION 2221–2226 (1992).

> options…suggesting why certain health-related values are more worthy and should be aspired to." Deliberation between the doctor and patient remains limited to "health-related values, that is, values that affect or are affected by the patient's disease and treatments; he or she recognizes that many elements of morality are unrelated to the patient's disease or treatment and beyond the scope of their professional relationship." The aim of the deliberation is moral persuasion, but not coercion, with the patient ultimately deciding on the appropriate validity and priority of these values in their life. Whereas the doctor is an advisor or counsellor in the interpretive model, in the deliberative model they serve as "a teacher or friend, engaging the patient in dialogue on what course of action would be best." The doctor indicates both what the patient could do and, in the context of their understanding of the patient's life and values, what he thinks the patient should do in terms of choice of intervention. The final decision still remains with the patient but is subject to greater persuasion and normative argumentation on the part of the doctor. This model conceives of patient autonomy as a tool for moral self-development; "the patient is empowered not simply to follow unexamined preferences or examined values, but to consider, through dialogue, alternative health-related values, their worthiness, and their implications for treatment."

A fifth model is mentioned in Emanuel and Emanuel's treatment of the doctor-patient relationship, the 'instrumental model', but quickly discarded on moral grounds. In the instrumental model the patient's values are given no importance; rather, the doctor takes a decision or convinces the patient to choose a particular course of treatment on the basis of external values such as social or scientific good. While rightly condemned on moral grounds, it should be noted that this model remains potentially relevant as a warning for the deployment of AI. In cases where AI is pursued not for the good of the patient, but rather for the sake of efficiency or cost savings, one could argue the doctor-patient relationship is instrumentalized. The influence of such external values on the doctor-patient relationship are elaborated below.

Each of these models of the doctor-patient relationship show varying degrees of respect to patient autonomy and moral self-development. The rights and values embedded in the Oviedo Convention provide some indication of the general acceptability of these models of the doctor-patient relationship. A paternalistic model would appear prone to violating the informed consent requirement set out in Article 5. A deliberative model would likewise appear to violate a specific aspect of the consent requirement expanded on in the Convention's Explanatory Report: a patient's consent should be based on "objective information" provided "in the absence of any pressure from anyone." The difficulty of providing objective information will be picked up again in the section entitled "Potential impact of AI on the doctor-patient relationship" in discussing transparency in AI-mediated clinical care.

Professional ethics in medicine

The Oviedo Convention explicitly calls for quality standards to be set by member states and professional societies in Article 4. But how does medicine as a profession set its own standards for clinical care and the doctor-patient relationship, and according to which goals or values? To this end, this section proposes a theoretical framework for understanding medicine as a self-governing profession. This framework aligns with many of the values prescribed in the Oviedo Convention; this aspect is further discussed in the section entitled "Potential impact of AI on the doctor-patient relationship".

An influential approach which prescribes ideal ends (and thus norms and internal goods) of medicine based upon virtue ethics has been advanced by Pellegrino and Thomasma.[125] Within this approach, based upon Alisdair MacIntyre's virtue ethics,[126] medicine can be considered a "moral practice"[127] with virtues describing character traits required of doctors in addition to the "medical scientific knowledge, practical skills and experience that ensures that the doctor does the right things with the right attitude in order to reach the goals of medicine."[128] Medicine is a moral practice by MacIntyre's definition because as a profession it self-governs, defines, and upholds internal standards of good medical care and accreditation processes to uphold these standards.[129]

The telos of a practice can be understood through critical examination of its internal goods or norms of evaluation; for medicine, these norms can be found in the doctor-patient relationship.[130] As seen in this relationship, "the ends of medicine are...the restoration or improvement of health and, more proximately, to heal, that is, to cure illness and disease or, when this is not possible, to care for and help the patient to live with residual pain, discomfort or disability."[131] The doctor-patient relationship, understood as a type of "healing relationship," is the primary mechanism through which these ends are realised.

Treating medicine as a moral practice with norms of good practice realised through a healing relationship is not to adapt an antiquated view of medicine as a paternalistic patient-provider relationship. Rather, the healing relationship involves both clinical interventions and information or services provided to patients for the sake of knowledge, empowerment or self-care. Even in modern clinical encounters with patients 'empowered' with democratised access to medical information, personal

[125] PELLEGRINO AND THOMASMA, *supra* note 120 at 52.
[126] MACINTYRE, *supra* note 122.
[127] PELLEGRINO AND THOMASMA, *supra* note 120.
[128] Petra Gelhaus, *The desired moral attitude of the physician: (I) empathy*, 15 MEDICINE, HEALTH CARE AND PHILOSOPHY 103–113, 104 (2012).
[129] PELLEGRINO AND THOMASMA, *supra* note 120; PAUL STARR, THE SOCIAL TRANSFORMATION OF AMERICAN MEDICINE (REVISED EDITION): THE RISE OF A SOVEREIGN PROFESSION AND THE MAKING OF A VAST INDUSTRY (2nd Revised ed. edition ed. 2017); General Medical Council, *Consent Guidance* (2008), http://www.gmc-uk.org/guidance/ethical_guidance/consent_guidance_index.asp; GENERAL MEDICAL COUNCIL, *supra* note 122.
[130] PELLEGRINO AND THOMASMA, *supra* note 120 at 52.
[131] *Id.* at 52–3.

values and lived experience with disease,[132] the doctor as an ideal-type 'role' requiring certain technical expertise and professional training is beyond question—the point of contention is rather whether this expertise should be deferred to without challenge.

Fiduciary duties and the healing relationship

Human rights principles regarding health and supportive rights enacted through policies such as the Charter of Fundamental Rights of the European Union reflect the moral and fiduciary duties of medicine as a profession. As discussed above, these obligations can be traced to the core aims or ends of medicine as a practice, and can be traced to many possible theoretical foundations, including human rights, care ethics and feminist ethics, and virtue ethics.

The remainder of this section focuses on an account of the healing relationship and medicine's fiduciary duties developed in the context of virtue ethics. A virtue-based approach emphasises the importance of treating the patient as a whole and promoting the patient's well-being through good practice. Standards are defined against goods such as compassion that "safeguards that the patient is not only seen as a number,"[133] contextual understanding of the patient's values, history and concerns, an "interest in the inner processes of the patient…an adequate skill in responding non-verbally and by skilful and sensitive dialogue,"[134] alongside technical skill in 'fixing' the patient's disorder or managing a persistent condition. With that said, these core aims are shared by many other approaches outside of virtue ethics. For example, approaches to care ethics and feminist ethics focus on related goods such as the caring role of the health professional, relationships and care responsibilities (in contrast to a focus on justice and rights),[135] tacit knowledge and context-sensitive care that responds to the interests and needs of patients as unique, socially embedded individuals, and power imbalances and coercion owing to the vulnerable position of the patient.

Several characteristics of the healing relationship create moral obligations on practitioners to protect the interests of patients.[136] Specifically, the relationship can be characterised by the following traits:

- **Vulnerability and Inequality** – Patients experience a loss of control to define and pursue personal goals, and may experience emotional stress, fear, worry, and anxiousness.[137] The immediate goal of life becomes the restoration of health and well-being by relieving or curing symptoms. An imbalanced

[132] Emanuel and Emanuel, *supra* note 123; Edgar, *supra* note 118.
[133] Petra Gelhaus, *The desired moral attitude of the physician: (II) compassion*, 15 MEDICINE, HEALTH CARE AND PHILOSOPHY 397–410, 405 (2012).
[134] Gelhaus, *supra* note 127 at 108.
[135] CAROL GILLIGAN, IN A DIFFERENT VOICE: PSYCHOLOGICAL THEORY AND WOMEN'S DEVELOPMENT (1993).
[136] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4; Schotsmans, Dierckx de Casterle, and Gastmans, *supra* note 120.
[137] PELLEGRINO AND THOMASMA, *supra* note 120; David B. Morris, *About suffering: Voice, genre, and moral community*, 125 DAEDALUS 25–45 (1996); Keith Bauer, *Cybermedicine and the moral integrity of the physician–patient relationship*, 6 ETHICS AND INFORMATION TECHNOLOGY 83–91 (2004); Deborah Lupton, *The digitally engaged patient: self-monitoring and self-care in the digital health era*, 11 SOCIAL THEORY & HEALTH 256–270, 263 (2013).

relationship is created in which the patient is compelled to seek the help of an individual with privileged medical expertise in the pursuit of a return to health. Doctors have an obligation to not use their expertise or privileged position of power to exploit the "vulnerable" patient.[138]

- **Fiduciary Nature** – The patient explicitly or tacitly places trust in a chosen doctor and reveal aspects of himself and his life to allow diagnosis and healing, surrendering some privacy in allowing "others access to personal information or [their] bodies."[139] Doctors have a moral obligation to make use of the information and access provided by the patient in a trusting relationship in the patient's best interests, and not for self-interest.[140]
- **Nature of Medical Decisions** – Medical decisions are a combination of technical and moral features. The doctor's diagnosis and treatment of the patient must be technically accurate to promote physical health.[141] However, decisions should also support the patient's moral well-being or autonomy as an entity with moral value, in the sense that the decision should match with the patient's values.[142]
- **Characteristics of Medical Knowledge** – Medical knowledge is non-proprietary. To ensure a sufficient quantity of health professionals, societies provide doctors with privileged knowledge and access to human bodies necessary to gain medical expertise and may limit recognition of practitioners of medicine to individuals thus trained. Doctors have a moral obligation to act as stewards to this knowledge, ensuring it is readily available to others, used ethically in the treatment of patients, and not purely for self-interest.[143]
- **Moral Complicity** – The doctor is the channel through which medical interventions flow to the patient, in the sense that the doctor must agree to each intervention carried out. In this position the doctor has a moral obligation to act as a gatekeeper, safeguarding the patient's well-being and acknowledging his complicity in any interventions carried out.[144]

These characteristics are not beyond question; for instance, the experience of illness as vulnerability and inequality can be criticised in that it only seems to apply to acute problems with potential cures.[145] Although the 'healing relationship' approach

---

[138] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6; GILLIGAN, *supra* note 134.
[139] BEAUCHAMP AND CHILDRESS, *supra* note 123 at 298.
[140] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4; Bauer, *supra* note 136; John Heritage et al., *Problems and Prospects in the Study of Physician-Patient Interaction: 30 Years of Research*, 32 ANNUAL REVIEW OF SOCIOLOGY 351–374, 355 (2006); O. Karnieli-Miller & Z. Eisikovits, *Physician as partner or salesman? Shared decision-making in real-time encounters*, 69 SOCIAL SCIENCE & MEDICINE 1–8, 2 (2009).
[141] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4.
[142] BEAUCHAMP AND CHILDRESS, *supra* note 123; Karnieli-Miller and Eisikovits, *supra* note 139.
[143] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4.
[144] *Id.* at 35–6, 42–4.
[145] MARTHA C. NUSSBAUM, FRONTIERS OF JUSTICE DISABILITY, NATIONALITY, SPECIES MEMBERSHIP (OIP): DISABILITY, NATIONALITY, SPECIES MEMBERSHIP (TANNER LECTURES ON HUMAN VALUES) (New Ed ed.

describes an idealistic model of the doctor-patient relationship (and thus, medicine itself), the underlying notion that being a doctor includes moral obligations to the patient is widely accepted.[146] The fundamental character of the medical relationship as one in which a patient in need seeks medical knowledge, expertise, or treatment is beyond question. In seeking out professional help, the patient is tacitly agreeing to reveal herself and private aspects of her life to the doctor with medical expertise in the pursuit of health. The relationship is an exchange of sensitive goods for improvements in quality of life which the patient is coerced through illness to engage in if a return to health is desired. Doctors are consulted not merely as 'encyclopaedias of knowledge', but rather as 'trusted' experts capable of subjective evaluation and understanding the patient as a socially embodied person with a history and values.[147]

Being a medical professional, or belonging to medicine understood as a formal profession, requires committing oneself to the moral obligations of the healing relationship.[148] Medicine can be considered a 'moral practice' in this context because its members form a community which shares a common goals and moral obligations,[149] meaning they are "guided by some shared source of morality—some fundamental rules, principles, or character traits that will define a moral life consistent with the ends, goals, and purposes of medicine".[150] Critically, this account contrasts the norms and obligations of individual practitioners with those of the institutions through which care is provided. Whereas the individual health professional's first obligation is to the patient, institutions have other (legitimate) interests concerning resourcing and quality of care across the institution as a whole. From a virtue ethics perspective, medical virtues and internal norms of good practice can help ensure the ends of medicine, and ultimately the obligations to individual patients incurred through the healing relationship, are met over time and resist erosion due to the corrupting influence of institutions and external goods.[151] For a discussion of specific virtues of good medical practice, see the Appendix.

Emergent challenges in the doctor-patient relationship

It could be argued that the healing relationship model is outdated, as "the notion of patients placing themselves under the care of a doctor and seeking their expert advice has moved to the concept of patients as producing health knowledges and as acquiring expert knowledge so as to manage their illness themselves."[152] This

2007); Barbara Page-Hanify, *Intellectual Handicap - Achievement of Potential*, 27 AUSTRALIAN OCCUPATIONAL THERAPY JOURNAL 53–60 (1980).

[146] BEAUCHAMP AND CHILDRESS, *supra* note 123; Andrew Edgar & Stephen Pattison, *Integrity and the moral complexity of professional practice*, 12 NURSING PHILOSOPHY 94–106 (2011); Gelhaus, *supra* note 127; Y. M. Barilan & M. Brusa, *Deliberation at the hub of medical education: beyond virtue ethics and codes of practice*, 16 MEDICINE, HEALTH CARE AND PHILOSOPHY 3–12 (2013).

[147] Emanuel and Emanuel, *supra* note 123 at 2225; Gelhaus, *supra* note 127 at 110.

[148] STARR, *supra* note 128.

[149] PELLEGRINO AND THOMASMA, *supra* note 120 at 3; Morris, *supra* note 136; Schotsmans, Dierckx de Casterle, and Gastmans, *supra* note 120.

[150] PELLEGRINO AND THOMASMA, *supra* note 120 at 3.

[151] *Id.* at 32.; MACINTYRE, *supra* note 122.

[152] Deborah Lupton, *M-health and health promotion: The digital cyborg and surveillance society*, 10 SOCIAL THEORY & HEALTH 229–244, 233 (2012).

characterisation of medicine suggests that the doctor-patient relationship has evolved and can seamlessly incorporate AI without altering the character of medical care.

As the practice of medicine changes in the face of emerging technologies, "something of the past is inevitably lost, not always for the worse."[153] Medicine has long been affected by advances in technology that disrupt the traditional one-to-one, face-to-face model of clinical care between doctor and patient. The Internet, for example, has empowered patients with greater access to medical information, but introduced risks owing to misleading or inaccurate information. Introducing new stakeholders into care relationships is not self-evidently problematic, but must be measured in terms of impact on the healing relationship and the ends of medicine; in other words, in the impact on patient care.

The healing relationship must be understood as an idealistic framework of the relationship between 'expert' doctors and 'vulnerable' patients. As an ideal, the model is not reflective of the 'empowered patient' model of care that has emerged in parallel over the past several decades.[154] Assuming modern medicine is characterised by 'empowered' patients eroding the privileged position of doctors as 'experts', trust cannot be assumed to exist whenever healing occurs.

However, the healing relationship describes the motivations of patients to seek professional care, or knowledge and technologies for self-care. Whether addressed through professional or self-directed care, the vulnerability of the patient is not eliminated. Similarly, the fiduciary duties created by this vulnerability do not change when diffused to different sources of expertise, be they medical professionals, databases of medical knowledge and advice, or other technologies and systems supporting self-care such as telemedicine or readily available medical information on the Internet.

Finding new ways to live up to the fiduciary duties of medicine in practice takes on renewed importance in this context and in the future deployment of AI in medicine. Pertinent questions have been asked, for example, about the validity and efficacy of medical knowledge available through internet portals. Furthermore, although medical information is increasingly available through other mediums, the role of expertise as an indication of fidelity to trust does not change.[155] Providers of low-quality medical advice, information or care can be criticised, regardless of format.

On this basis, the healing relationship model can be understood as a description of the moral character and obligations of medical practice, traditionally embodied by health practitioners but increasingly diffused across various platforms and persons, including web portals, consumer device developers, providers of wellness services, and others. Even if modern medicine has moved beyond the single doctor-patient model described in the healing relationship, the obligations of this relationship have not disappeared. Rather, the diffusion and displacement of these obligations by new technological actors in medicine is a cause for concern in considering how best to

[153] PELLEGRINO AND THOMASMA, *supra* note 120 at 32.
[154] Emanuel and Emanuel, *supra* note 123.
[155] *Id.*

govern the introduction of AI in medicine. Our notion of the healing relationship could, of course, be revised to give primacy to patient autonomy above all else. However, doing so risks reducing the doctor to a mere service-provider, incapable of exercising the full range of medical virtues and practice-internal norms.

When evaluating the impact of AI and algorithmic technologies on the doctor-patient relationship, choice of metric is key. If measured solely in terms of cost-benefits, or utility, the justification for AI mediation and augmentation of care is straightforward. However, while algorithmic technologies may allow for a greater number of patients to be treated more efficiently or at lower cost, their usage can simultaneously undermine non-mechanical dimensions of care. A distinction can be drawn between those effects of algorithmic systems (and components of utility) which contribute to the good of the patient or medicine as a practice governed by well-established internal norms and codes of conduct, and those which contribute to the good of medical institutions and healthcare services.

The moral complicity that characterises the doctor-patient relationship, wherein treatment is ideally guided by the professional's contextually and historically aware assessment of a patient's condition, cannot be easily replicated in interactions with AI systems. The role of the patient, the factors that lead people to seek medical attention, and the patient's vulnerability are not changed by the introduction of AI as a mediator or augmenter of medical care. Rather, what changes is the means of care delivery, how it can be provided, and by whom. The shift of expertise and care responsibilities to AI systems can be disruptive in many ways, which are explored in the section entitled "Potential impact of AI on the doctor-patient relationship".

6 POTENTIAL IMPACT OF AI ON THE DOCTOR-PATIENT RELATIONSHIP

AI promises a variety of opportunities, benefits, and risks for the practice of medicine. Drawing on the framework of ethical challenges facing AI and policy context developed in the sections entitled "Background and context", "Overview of AI applications in medicine", and "Theoretical framework of the doctor-patient relationship", this section identifies six potential impacts of AI on the doctor-patient relationship.

Inequality in access to high quality healthcare

As an emerging technology the deployment of AI systems will not be immediate or universal across all member states or healthcare systems. Deployment across institutions and regions will inevitably be inconsistent in terms of scale, speed, and prioritisation. Telemedicine systems, for instance, are well suited to providing access to care in remote or inaccessible places, or where shortages exist in healthcare workers or specialists.[156] This promises to fill gaps in healthcare coverage but not necessarily with care of equivalent quality to traditional face-to-face care. Impact on the doctor-patient relationship in the near term may therefore be much greater in areas suffering from existing staffing shortages or new shortages owing to the COVID-19 pandemic. The quality and degree of this impact remains to be seen.

The unavoidable variability in deployment of AI raises the possibility that geographical bias in performance and inequalities in access to high quality care will be created through the usage of AI systems. This cuts both ways. If AI systems raise the quality of care, for example by providing more accurate or efficient diagnosis, expanded access to care, or through the development of new pharmaceutical and therapeutic interventions, then patients served by 'early adopter' regions or health institutions will benefit before others. AI systems may also be used to free up clinicians from menial, labour intensive tasks such as data entry and thus provide more time with patients than was previously possible.[157]

However, these benefits are not foregone conclusions. The impact of AI on clinical care and the doctor-patient relationship remains uncertain and will certainly vary by application and use case. AI systems may prove to be more efficient than human care, but also provide lower quality care featuring fewer face-to-face interactions. In many areas AI is seen as a promising means to cut costs, reduce waiting times, or fill existing gaps in coverage where access to health professionals and institutions is limited.[158] Patients in early adopter areas will at a minimum receive a different type of care which

[156] World Health Organization, *supra* note 1.
[157] *Id.* at 8.
[158] Department of Health, *Innovation Health and Wealth: Accelerating Adoption and Diffusion in the NHS* (2011); DEPARTMENT OF HEALTH, EQUITY AND EXCELLENCE: LIBERATING THE NHS. (2010).

may not be of the same quality as traditional care provided by human health professionals.

The inconsistent rollout of AI systems with uncertain impacts on access and care quality poses a risk of creating new health inequalities in member states. It may prove to be the case that regions that have historically faced unequal access or lower quality care are seen as key test beds for AI-mediated care. Patients in these areas may have better access to AI systems, such as chatbots or telemedicine, but continue to face limited access to human care or face-to-face clinical encounters. The likelihood of this risk depends largely on the strategic role given to AI systems. If they are treated as a potential replacement for face-to-face care, rather than as a means to free up clinicians' time greater inequality in access to human care seems inevitable.

Article 4 of the Oviedo Convention addresses care provided by healthcare professionals bound by professional standards. It remains unclear whether developers, manufacturers, and service providers for AI systems will be bound by the same professional standards. The Convention's Explanatory Report raises this question indirectly, noting that "from the term 'professional standards' it follows that [Article 4] does not concern persons other than health care professionals called upon to perform medical acts, for example in an emergency." Can a chatbot designed for initial triage of patients be considered a "person" performing a "medical act"?[159] If not, how can the involvement of an appropriately bounded healthcare professional be guaranteed?

Any reduction in oversight or clinical care by health professions caused by the rollout of AI systems could thus potentially be viewed as a violation of Article 4. In particular, care models that incorporate chat bots or other artificial agents designed to provide care or support directly to patients would seem to pose this risk. Careful consideration must be given to the role played by healthcare professions bound by professional standards when incorporating AI systems that interact directly with patients.

Transparency to health professionals and patients

AI challenges our notions of accountability in both familiar and new ways. Systems increasingly trusted to help make life-changing decisions and recommendations have their foundation in our technological past, but they are digital, distributed, and often imperceptible. When important decisions are taken which affect the livelihood and well-being of people, one expects that their rationale or reasons can be understood.

This expectation is reflected in Article 5 of the Oviedo Convention which reaffirms the right to informed consent for patients prior to being subject to medical interventions or research. As detailed above, the Convention's Explanatory Report specifies a non-comprehensive list of information to be provided. An overarching requirement is that the information must be provided to patients in an easily understandable way to ensure it can meaningfully inform their decisions. Traditionally, this would impose requirements on how health professionals explain their decisions and

[159] Korngiebel and Mooney, *supra* note 95 at 3.

recommendations to patients. In cases where AI systems provide some form of clinical expertise, for example by recommending a particular diagnosis or interpreting scans, this requirement to explain one's decision-making would seemingly be transferred from doctor to AI system, or at least to manufacturer of AI system.

The difficulty of explaining how AI systems turn inputs into outputs poses a fundamental epistemological challenge for informed consent. Aside from the patient's capacity to understand the functionality of AI systems, in many cases patients simply do not have sufficient levels awareness to make free and informed consent possible. AI systems use unprecedented volumes of data to make their decisions, and interpret these data using complex statistical techniques, both of which add to increase the difficulty and effort required to remain aware of the full scope of data processing informing one's diagnosis and treatment.[160]

In practice, transparency requirements in the service of informed consent can be borne out in several ways. Assuming doctors remain as the primary point of care for patients, the doctor can be seen as a mediator between the patient and the AI system. In this mediation model, the doctor can be the recipient of an explanation from the AI system and then act as a 'translator' for the patient, translating the system's explanation into a meaningful and easily understandable format. Where doctors do not act as mediators, for example where chatbots provide diagnosis or triage directly to patients, AI systems may then be expected to explain their decision-making directly to patients.

Both models pose challenges in explaining complex 'black box' behaviours to expert or non-expert users. At a minimum, AI systems interacting directly with patients should self-identify as an artificial system. Whether any usage of AI systems in care should be disclosed to patients by clinicians and healthcare institutions is a more difficult question.[161]

A commonly cited concern with AI used for operational purposes by hospitals, including risk stratification and discharge, planning tools is a failure to inform patients about the usage of AI in their care.[162]

On the one hand, health professionals routinely consult many sources of information in diagnosing and treating patients, such as models, charts, X-rays, etc., that they would not disclose or proactively discuss as part of informed consent. On the other hand, AI systems which effectively provide artificial clinical expertise, for instance by interpreting scans and recommending a classification of abnormalities, may be a qualitatively different type of information than sources that traditionally factor into clinical decision-making.

Nonetheless, in practice AI systems used to support clinical care and stratify risk among patients are often treated as purely operational rather than clinical applications. According to many health institutions they are used to improve the quality and

[160] COUNCIL OF EUROPE, *supra* note 2.
[161] I. Glenn Cohen, *Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? Symposium: Law and the Nation's Health*, 108 GEO. L.J. 1425–1470 (2019); Robbins and Brodwin, *supra* note 5.
[162] Cohen, *supra* note 160; Robbins and Brodwin, *supra* note 5.

efficiency of care, not to inform clinical decision-making. In this regard, they can be considered equivalent to other administrative systems used in hospitals that handle patient data but not for their immediate care.[163] Of course, not all health institutions treat AI risk prediction systems as purely operational; in some cases, patients are asked to explicitly consent to the usage of an AI system designed to identify patients at risk of death in the next 48 hours.[164] Recommendations concerning disclosure of the usage of AI systems will be returned to in the Section entitled "Public register of medical AI systems for transparency".

Independent of the question of whether particular AI applications should be classified as clinical or operational/administrative, there are pertinent questions concerning the intelligibility of 'black box' systems at a more fundamental level. Compared to human and organisational decision-making, AI poses a unique challenge. The internal state of a trained machine learning model can consist of millions of features connected in a complex web of dependent behaviours. Conveying this internal state and dependencies in a human comprehensible way is extremely challenging.[165] How AI systems make decisions may thus be too complex for human beings to thoroughly understand their full decision-making criteria or rationale.

Assuming the transparency requirement underlying informed consent is a key value in the AI-mediated doctor-patient relationship, the challenge of opacity raises a question: how should AI systems explain themselves to doctors and patients? We can begin to unpack this question by examining the different types of questions, notably we may ask about AI systems to make them understandable:

- **How does an AI system or model function? How was a specific output produced by an AI system?** These are questions of interpretability. Questions of interpretability address the internal functionality and external behaviour of an AI system. A fully interpretable model is one which is human comprehensible, meaning a human can understand the full set of causes of a given output.[166] Poorly interpretable models 'are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs'.[167] Interpretability can also be defined in terms of the predictability of the model; a model is interpretable if a well-informed person could consistently predict its outputs and behaviours.[168] Questions of model behaviour

[163] Robbins and Brodwin, *supra* note 5.
[164] *Id.*
[165] Jenna Burrell, *How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOCIETY (2016); Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV:1606.03490 [CS, STAT] (2016), http://arxiv.org/abs/1606.03490 (last visited Oct 15, 2016).
[166] Paulo JG Lisboa, *Interpretability in Machine Learning–Principles and Practice*, *in* FUZZY LOGIC AND APPLICATIONS 15–21 (2013), http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2 (last visited Dec 19, 2015); Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 267 ARTIFICIAL INTELLIGENCE 1–38 (2019).
[167] Burrell, *supra* note 164 at 1.
[168] Been Kim, Rajiv Khanna & Oluwasanmi O. Koyejo, *Examples are not enough, learn to criticize! criticism for interpretability*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 2280–2288 (2016).

narrowly address how a particular output or behaviour of the model occurred.[169] However, model behaviour can also be broadly interpreted to include effects on reliant institutions and users and their AI-influenced decisions, for example how a doctor's diagnosis was influenced by an expert system's recommendation, are also relevant.[170]

- **How was an AI system designed and tested? How is it governed?** These are questions of transparency. Unlike interpretability, transparency does not address the functionality or behaviour of the AI system itself, but rather the processes involved in its design, development, testing, deployment, and regulation. Transparency principally requires information about the institutions and people that create and use AI systems, as well as the regulatory and governance structures that control both the institutions and systems. Here, interpretability play a supplementary but supportive role. Interpretable models or explanations of specific decisions taken by a system may, for example, be needed for regulators to effectively audit AI and ensure regulatory requirements are being met in each context of use.
- **What information is required to investigate the behaviour of AI systems?** This is a question of traceability. To audit the behaviour of AI systems, certain evidence is needed, which can include 'data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used'.[171] This data needs to be consistently recorded as the system operates for effective governance to be feasible. Traceability is thus a fundamental requirement for post hoc auditing and explanations of model behaviour; without the right data, explanations cannot be computed after a model has produced a decision or other output.[172]

Answers to each of these questions may be necessary to achieve informed consent in AI-mediated care. This is not to say both patients and health professions require answers to each question; rather, it may be the case that certain questions are better directed towards one or the other. For example, patients may be most immediately interested in questions concerning how their specific case was decided, or a diagnosis or recommendation reached.[173] Questions concerning how AI systems have been designed and tested, and how they are secured and validated over time, may be more immediately relevant to health professionals and administrators who must assess a system's trustworthiness in terms of integrating it into existing clinical and operational

[169] The degree to which the reasons for specific model behaviours can be explained is sometimes referred to as the *explainability* of a model. Here it is treated as one component of *interpretability* alongside intrinsic model comprehensibility.
[170] HIGH LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, *Ethics Guidelines for Trustworthy AI* (2019).
[171] *Id.*
[172] Mittelstadt et al., *supra* note 17.
[173] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 3 HARVARD JOURNAL OF LAW & TECHNOLOGY 841–887 (2018).

decision-making pathways.[174] As suggested in the section entitled "Theoretical framework of the doctor-patient relationship", the informed consent ideal is one component of the doctor-patient relationship requiring discussion between patients and health professionals of possible treatment options, values, and the like. Directing explanation types to the parties best equipped to understand them, or most immediately interested in them, need not undermine ideals of transparency or informed consent, but rather can be seen as a facilitator of meaningful dialogue between patient and doctor about options in AI-mediated care.

Risk of social bias in AI systems

As discussed in the section entitled "Common ethical challenges in AI", AI systems are inevitably biased in some respect. Many biases arise due to technical reasons, such as a mismatch between training and testing environments.[175] System developers and manufacturers inevitably design systems that reflect their values or relevant regulatory requirements; this can also be treated as a type of bias which will vary between manufacturers and member states.[176] However, in AI systems biased and unfair decision-making often occurs not for technical or regulatory reasons, but rather reflect underlying social biases and inequalities.[177]

These types of social biases are concerning for several reasons.

- First, they may undermine the accuracy of models across different populations or demographic groups. Many biases can be traced to datasets that are not representative of the population targeted by a system. In medicine, there are crucial data gaps that can be filled but to date are not due to limitations on resources, access, or motivation. Clinical trials and health studies are predominantly undertaken on white male subjects meaning results are less likely to apply to women and people of colour.[178] A serious and dangerous data gap exists because many clinical models treat women as "little men"[179] and thus do not account for biological differences.[180] For example, different percentage of body fat, thinner skin, different hormone levels and compositions, changing hormone levels throughout the menstrual cycle, changing hormone levels

[174] COUNCIL OF EUROPE, *supra* note 2.
[175] Friedman and Nissenbaum, *supra* note 41; Wachter, Mittelstadt, and Russell, *supra* note 47.
[176] COUNCIL OF EUROPE, *supra* note 2.
[177] *Id.*; Wachter, Mittelstadt, and Russell, *supra* note 47.
[178] CAROLINE CRIADO PEREZ, INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN 115–116 (2019); on how to adress bias in the medical setting see Timo Minssen et al., *Regulatory responses to medical machine learning*, JOURNAL OF LAW AND THE BIOSCIENCES (2020); and Mirjam Pot, Wanda Spahl & Barbara Prainsack, *The Gender of Biomedical Data: Challenges for Personalised and Precision Medicine*, 9 SOMATECHNICS 170–187 (2019).
[179] ANGELA SAINI, INFERIOR: HOW SCIENCE GOT WOMEN WRONG AND THE NEW RESEARCH THAT'S REWRITING THE STORY 59 (2017).
[180] PEREZ, *supra* note 177 at 116 One of the reasons why this is not done is because it is more complex (e.g. fluctuating hormone levels during the menstrual cycle), risky (e.g. female participants could be pregnant), time and resource intensive to study women. SAINI, *supra* note 178 at 58.

before puberty and after menopause are factors that affect how well drugs work or how much we are affected by toxins or environmental impacts.[181]

- Second, social biases can lead to unequal distribution of outcomes across populations or protected demographic groups. Inequality of this type is particularly severe in the context of medicine which affects fundamental goods: "any bias in the functioning of an algorithm could lead to inadequate prescriptions of treatment and subject entire population groups to unwarranted risks that may threaten not only rights but also lives."[182] Large segments of Western societies currently face significant prejudice and inequality which are captured in historical decisions and can influence the training of future systems. Historical trends in decision-making have led to diminished and unequal access to opportunities and outcomes among certain groups.[183] Without intervention, these pre-existing patterns in access to opportunities and resources in society will be learned and reinforced by AI systems.

As discussed, Article 14 of the ECHR prohibits discrimination. Equality is a key value underlying human rights. However, achieving substantive equality or a 'level playing field' in practice is extremely difficult. With regards to AI, dataset bias and feedback loops are key challenges to ensure systems do not exacerbate existing inequalities and create new forms of discrimination that would run counter to Article 14. The Parliamentary Assembly of the Council of Europe has recognised the risk of bias in this respect, noting that "Council of Europe member states should participate more actively in the development of AI applications for health care services, or at least provide some sort of sovereign screening and authorisations for their deployment. States' involvement would also help to ensure that such applications are fed with sufficient, unbiased and well protected data."[184]

Concerning dataset bias, conceiving of bias solely as a property of datasets is insufficient to achieve substantive equality in practice.[185] Assuming it is possible to create a dataset that perfectly captures existing biases and inequalities in society, training a model with this dataset would do nothing to correct the inequalities captured by it. Rather, such assurances can only be provided by also examining, testing for, and perhaps correcting biases in the trained AI system and its outputs.

With regards to feedback loops, reinforcing existing biases in society that have been learned by an AI system can make matters substantively worse for already disadvantaged groups. However, simply avoiding reinforcement of existing biases and inequalities, or ensuring AI systems do not make the status quo worse, does not achieve substantive equality in practice.[186] Rather, this requires critically examining the acceptability of existing inequalities and taking steps to positively improve the situation of disadvantaged groups. Likewise, AI systems can create novel forms of

[181] SAINI, *supra* note 178 at 62; PEREZ, *supra* note 177 at 116.
[182] COUNCIL OF EUROPE, *supra* note 2.
[183] See for example ANGELA Y. DAVIS, WOMEN, RACE, & CLASS (2011).
[184] COUNCIL OF EUROPE, *supra* note 2.
[185] Wachter, Mittelstadt, and Russell, *supra* note 47.
[186] *Id.*

discrimination rather than simply reinforcing existing forms of bias and inequality.[187] Both the need for critical positive action and the possibility of novel forms of discrimination fuelled by AI need to be accounted for in deploying AI in medicine.

Detecting biases in AI systems is not straightforward. Biased decision-making rules can be hidden in 'black box' models. Other biases can be detected by examining the outputs of AI systems for unequal distributions across demographic groups or relevant populations. However, accessing the full range of decisions or outputs of a system is not necessarily straightforward, at a minimum due to data protection standards; "certain restrictions on the use of personal health data may disable essential data linkages and induce distortions, if not errors, in AI-driven analysis."[188] At a minimum, this suggests that simply anonymising health data may not be an adequate solution to mitigate biases or correct their downstream effects. Even where decision sets are accessible, demographic data may not exist for the relevant populations meaning bias testing cannot measure distribution across relevant legally protected groups.[189]

These various challenges of social bias, discrimination, and inequality suggest health professionals and institutions face a difficult task in ensuring their usage of AI systems does not further existing inequalities and create new forms of discrimination. Combatting social bias is a multifaceted challenge which must include robust bias detection and testing standards, high quality collection and curation standards for training and testing datasets, and individual-level testing to ensure patient outcomes and recommendations are not predominantly determined by legally protected characteristics.[190] Failing to implement robust bias testing standards risks further exacerbating inequalities in AI-driven care and undermining the trustworthiness of AI-mediated care. These risks are particularly acute in the context of existing inequalities in access to high-quality care where the deployment of AI may be accelerated for the sake of efficiency and resource allocation rather than purely clinical considerations.

Dilution of the patient's account of well-being

Traditionally, clinical care and the doctor-patient relationship are ideally informed by the doctor's contextual, historically aware assessment of a patient's condition. This type of care cannot be easily replicated in technologically-mediated care. Data representations of the patient necessarily restrict the doctor's understanding of the patient's case to measured features. This can present a problem when clinical assessments increasingly rely on data representations, constructed for example by remote monitoring technologies, or other data not collected in face-to-face encounters. Data representations of patients can come to be seen as an 'objective' measure of health and well-being, reducing the importance of contextual factors of health or the view of the patient as a socially embodied person. Data representations can create a

[187] Wachter, Mittelstadt, and Russell, *supra* note 47.
[188] COUNCIL OF EUROPE, *supra* note 2.
[189] Wachter, Mittelstadt, and Russell, *supra* note 47; Sandvig et al., *supra* note 79; Brent Mittelstadt, *Automation, Algorithms, and Politics| Auditing for Transparency in Content Personalization Systems*, 10 INTERNATIONAL JOURNAL OF COMMUNICATION 12 (2016).
[190] Wachter, Mittelstadt, and Russell, *supra* note 47; Wachter, Mittelstadt, and Russell, *supra* note 47; Matt J. Kusner et al., *Counterfactual Fairness* (2017).

'veneer of certainty', in which 'objective' monitoring data is taken to represent a true representation of the patient's situation, losing sight of the patient's interpersonal context and other tacit knowledge.[191]

Medical professionals face this difficulty when attempting to incorporate AI systems into care routines. The amount and complexity of data and technologically derived recommendations about a patient's condition makes it difficult to identify when important contextual information is missing. Reliance upon data collected by 'health apps' or monitoring technologies (e.g., smart watches) as a primary source of information about a patient's health, for example, can result in ignorance of aspects of the patient's health that cannot easily be monitored. This includes essential elements of mental health and well-being such as the patient's social, mental, and emotional states. 'Decontextualisation' of the patient's condition can occur as a result, wherein the patient loses some control over how her condition is presented and understood by clinicians and carers.[192]

All of these possibilities suggest the encounters through which the basic trust necessary for a doctor-patient relationship is traditionally developed may be inhibited by technological mediation. Technologies which inhibit communication of "psychological signals and emotions" can impede the doctor's knowledge of the patient's condition, undermining "the establishment of a trusting and healing doctor-patient relationship."[193] Care providers may be less able to demonstrate understanding, compassion, and other desirable traits found within 'good' medical interactions in addition to applying their knowledge of medicine to the patient's case. As a mediator placed between the doctor and patient, AI systems change the dependencies between clinicians and patients by turning some degree of the patient's ongoing care over to a technological system. This can increase the distance between health professionals and patients thereby suggesting a loss of opportunities to develop tacit understanding of the patient's health and well-being.[194]

Risk of automation bias, de-skilling, and displaced liability

As discussed in the section entitled "Common ethical challenges in AI", the introduction of AI systems into clinical care poses a risk of automation bias, according to which clinicians may trust the outputs or recommendations of AI systems not due to proven clinical efficacy, but rather on the basis of their perceived objectivity, accuracy, or complexity.[195] Any deployment of AI systems designed to augment human decision-making with recommendations, warnings, or similar interventions runs the risk of introducing automation bias. Empirical work on the phenomenon is somewhat nascent, but one recent study showed how even expert decision-makers can be prone to automation bias over time for problematic reasons (e.g., the cost of an AI system

[191] Mark Coeckelbergh, *E-care as craftsmanship: virtuous work, skilled engagement, and information technology in health care*, 16 MEDICINE, HEALTH CARE AND PHILOSOPHY 807–816 (2013).
[192] Mittelstadt et al., *supra* note 3.
[193] Bauer, *supra* note 136 at 84.
[194] Coeckelbergh, *supra* note 190.
[195] Zarsky, *supra* note 31 at 121.

as a proxy for accuracy or equality).[196] The Council of Europe has clearly recognised the risk of automation bias in calling for guarantees that "AI-driven health applications do not replace human judgement completely and that thus enabled decisions in professional health care are always validated by adequately trained health professionals."[197]

Reliance on AI systems as clinical care providers or expert diagnostic systems can inhibit the development of skills, professional communities, norms of 'good practice' within medicine. This phenomenon is referred to as 'de-skilling',[198] and runs counter to what the WHO has referred to as 'human-centred AI' which supports and augments human expertise and skill development, rather than undermining or replacing them.[199] Medical professionals develop virtues or norms of good practice through their experiences of practicing medicine. To define norms, practitioners can draw on practical wisdom developed through their experience. Members of the medical profession form a community which shares common goals and moral obligations.[200] The virtues or internal norms of a practice help ensure its ends are met over time by combating the influence of institutions and external goods. The development, maintenance, and application of these norms can be displaced through technological mediation of care.

It follows that the development, maintenance, and application of internal norms necessary to meet moral obligations to patients can be undermined when care is technologically mediated, and thus provided in part by non-professional individuals and institutions. A potential exists for algorithmic systems to displace responsibilities traditionally fulfilled by medical professionals, while providing more efficient or 'better' care measured solely in terms of cost-benefit. To prevent the erosion of holistically good, not merely technically 'efficient', medical care, these moral obligations to benefit and respect patients in the first instance need to be taken seriously by new care and services providers that are not part of traditional medical communities. In other words, a gap in professional skills and accountability can be created by AI-mediated care.

De-skilling and automation bias also pose risks directly to patients. One function of human clinical expertise is to protect the interests and safety of patients. Risks to safety come from a variety of sources, including "malicious attacks on software, unethical system design or unintended system failure, loss of human control and the "exercise of digital power without responsibility" that can lead to tangible harm to human health, property and the environment."[201]

If this human expertise is eroded through de-skilling or displaced through automation bias, testing and evidence of clinical efficacy must fill the gap to ensure patient safety. A similar trade-off exists in relation to opacity and accuracy; some scholars have

---

[196] Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853 (2019).
[197] COUNCIL OF EUROPE, *supra* note 2.
[198] *Id.*; Coeckelbergh, *supra* note 190.
[199] World Health Organization, *supra* note 1.
[200] MACINTYRE, *supra* note 122.
[201] COUNCIL OF EUROPE, *supra* note 2; COUNCIL OF EUROPE, *Responsibility and AI* (2019), https://rm.coe.int/responsability-and-ai-en/168097d9c5.

argued that medical AI systems do not necessarily need to be explainable if their accuracy and clinical efficacy can be reliably validated.[202] In both cases the protection of vital patient interests, or the fiduciary obligations typically shouldered by health professionals, are transferred to providers of AI systems or the systems themselves.

As a result, to continue to ensure patient safety and replace the protection offered by human clinical expertise, robust testing and validation standards should be an essential pre-deployment requirement for AI systems in clinical care contexts. These standards should also address complementary non-clinical aspects of safety such as cybersecurity, malfunctioning and resilience.[203] While a seemingly obvious conclusion, the existence of such requirements and evidence meeting them cannot be taken for granted. As discussed in the section entitled "Overview of AI technologies in medicine", evidence of clinical efficacy does not yet exist for many AI applications in healthcare, which has justifiably proven a barrier to widespread deployment.

A related but equally important topic concerns liability for malfunctioning and other harmful effects of AI. As discussed in the section entitled "Overview of AI technologies in medicine", distributed responsibility is both a morally and legally difficult challenge. The Parliamentary Assembly of the Council of Europe has recognised the need to clarify the liability of stakeholders in AI including "developers to regulatory authorities, intermediaries and users (including public authorities, health-care professionals, patients and the general public)." Member states of the Council of Europe are called on to "elaborate a legal framework for clarifying the liability of stakeholders for the design, deployment, maintenance and use of health-related AI applications (including implantable and wearable medical devices) in the national and pan-European context, redefine stakeholder responsibility for risks and harms from such applications and ensure that governance structures and law enforcement mechanisms are in place to guarantee the implementation of this legal framework."[204] A 2019 report from the Council of Europe Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) explored the specific challenges of liability and responsibility gaps in AI in much greater detail than is possible here.[205]

Impact on the right to privacy

AI poses several unique challenges to the human right to privacy and complementary data protection regulations. As discussed in the section entitled "The Oviedo Convention and human rights principles regarding health", the Council of Europe is currently in the processing of ratifying amendments to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108 and CETS No. 223). These additional rights seek to provide individuals with greater transparency and control over automated forms of data processing. These

[202] Boris Babic et al., *Beware explanations from AI in health care*, 373 SCIENCE 284–286 (2021).
[203] COUNCIL OF EUROPE, *supra* note 2.
[204] *Id.*
[205] COUNCIL OF EUROPE, *supra* note 200.

rights will undoubtedly provide valuable protection for patients across a variety of use cases of medical AI.

One distinct challenge unique to AI worth further consideration concerns the usage of patient data for training and testing AI systems. Confidentiality in the doctor-patient relationship is a key value to protect the human right to privacy. At the same time, greater development, deployment, and reliance on AI systems in care may create a greater need to create or curate high-quality real-world patient datasets to train and test systems. Innovation can threaten privacy and confidentiality in two ways. First, there may be a greater pressure to re-purpose and grant third party access to (deidentified) patient data and electronic health records to test and develop AI systems.

Second, clinicians may be encouraged to prescribe additional tests and analysis not for their clinical value but rather due to their utility for training or testing AI systems. This has implications both in terms of rising costs for healthcare but also exposure of patients to unnecessary risks of data leakage or other breaches of privacy. The Oviedo Convention sets out a specific application of the right to privacy (Article 8 ECHR) which recognises the particularly sensitive nature of personal health information and sets out a duty of confidentiality for health care professionals. Any generation of data with questionable clinical value or clearly motivated by its utility solely for the testing or development of AI systems would seemingly violate the Convention's specification of the right to privacy.

As this suggests, where a legitimate need exists for real-world data to test and train AI systems, interests in innovation and care efficiency or quality must be balanced with the patient's individual interests in privacy and confidentiality. Failing to strike this balance risks undermining trust between patients and care providers. Trust would be lost not owing to a failure to use AI appropriately in individual clinical encounters, but rather due to an institutional failure to protect patient interests in privacy and confidentiality at an institutional level. At a minimum, any re-purposing of patient health records for training and testing AI systems should be subject to sufficient deidentification and privacy enhancing techniques such as differential privacy (which introduces noise to prevent identification of a particular person in the dataset).[206]

[206] Cynthia Dwork, *Differential Privacy*, *in* AUTOMATA, LANGUAGES AND PROGRAMMING 1–12 (Michele Bugliesi et al. eds., 2006), http://link.springer.com/chapter/10.1007/11787006_1 (last visited Apr 4, 2016); Paul Ohm, *Broken promises of privacy: Responding to the surprising failure of anonymization*, 57 UCLA LAW REVIEW 1701 (2010).

7 RECOMMENDATIONS FOR COMMON ETHICAL STANDARDS FOR TRUSTWORTHY AI

The preceding discussion in the section entitled "Potential impact of AI on the doctor-patient relationship" concluded that ethical standards need to be developed around transparency, bias, confidentiality, and clinical efficacy to protect patient interests in informed consent, equality, privacy, and safety. Together, such standards could serve as the basis for deployments of AI in healthcare that help rather than hinder the trusting relationship between doctors and patients. These standards can address both how systems are designed and tested prior to deployment, as well as how they are implemented in clinical care routines and institutional decision-making processes.

The Oviedo Convention acts as a minimum standard for the protection of human rights which requires translation into domestic law. On this basis, there is an opportunity to make specific, positive recommendations concerning the standard of care to be met in AI-mediated healthcare. These recommendations must not interfere with the exercise of national sovereignty in standard setting through domestic law and professional bodies as detailed in Article 4 of the Oviedo Convention. However, it is also possible to set standards which do not interfere with Article 4 and can be considered directly enforceable. Specifically, as noted by Andorno:

> "The common standards set up by the Council of Europe will mainly operate through the intermediation of States. This does not exclude of course that some norms contained in the Convention may have self-executing effect in the internal law of the States having ratified it. This is the case, for instance, of some norms concerning individual rights such as the right to information, the requirement of informed consent, and the right not to be discriminated on grounds of genetic features. Prohibition norms can also be considered to have immediate efficacy, but in the absence of legal sanctions, whose determination corresponds to each State (Article 25), their efficacy is restricted to civil and administrative remedies."

Where AI can be observed to have a clear impact on rights and protections set out in the Oviedo Convention, it is appropriate for the Council of Europe to introduce binding recommendations and requirements for signatories concerning how AI is deployed and governed. Recommendations should focus on a higher positive standard of care with regards to the doctor-patient relationship to ensure it is not unduly disrupted or by the introduction of AI in care settings. Of course, such standards should be supportive to a degree of local interpretation around key normative issues like acceptable degrees of automation bias, acceptable trade-offs between outcomes between patient groups, and similar areas influenced by local norms.

The following example recommendations detail possible essential requirements and recommendations for an intelligibility standard that aims to protect informed consent in AI-mediated care, a transparency standard for public intelligibility, and a standard for collection of sensitive data for purposes of bias testing. Each should be treated as an example of the type of recommendation that can be drawn from the preceding discussion of the potential ethical impacts of AI on the doctor-patient relationship.

Intelligibility requirements for informed consent

According to the Explanatory Report, Article 5 of the Oviedo Convention contains an incomplete list of information that should be shared as part of an informed consent process. As this list is incomplete, the Council of Europe could set standards for what and how information about the recommendation of an AI system concerning a patient's diagnosis and treatment should be communicated to the patient. Given the traditional role of the doctor in sharing and discussing this type of information in clinical encounters, these standards should likewise address the doctor's role in explaining AI recommendations to patients and how AI systems can be designed to support the doctor in this role.

Several concepts are common across the questions and goods that motivate interpretability in AI. Interpretability methods seek to explain the functionality or behaviour of the 'black box' machine learning models that are a key component of AI decision-making systems. Trained machine learning models are 'black boxes' when they are not comprehensible to human observers because their internals and rationale are unknown or inaccessible to the observer, or known but uninterpretable due to their complexity.[207] Interpretability in the narrow sense used here refers to the capacity to understand the functionality and meaning of a given phenomenon, in this case a trained machine learning model and its outputs, and to explain it in human understandable terms.[208]

'Explanation' is likewise a key concept in AI interpretability. Generically, explanations in AI relate 'the feature values of an instance to its model prediction in a humanly understandable way'.[209] This rough definition hides significant nuance. The term captures a multitude of ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders.[210]

To understand how 'explanation' can be operationalised in medicine, two key distinctions are relevant:

---

[207] Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUT. SURV. 93:1-93:42 (2018); INFORMATION COMMISSIONER'S OFFICE & THE ALAN TURING INSTITUTE, *Explaining decisions made with AI* (2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[208] Finale Doshi-Velez & Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, ARXIV:1702.08608 [CS, STAT] (2017), http://arxiv.org/abs/1702.08608 (last visited Sep 22, 2017).

[209] CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING 31 (2020), https://christophm.github.io/interpretable-ml-book/ (last visited Jan 31, 2019).

[210] Lipton, *supra* note 164; Miller, *supra* note 165.

- First, methods can be distinguished in terms of what it is they seek to explain. Explanations of model functionality address the general logic the model follows in producing outputs from input data. Explanations of model behaviour, in contrast, seek to explain how or why a particular behaviour exhibited by the model occurred, for example how or why a particular output was produced from a particular input. Explanations of model functionality aim to explain what is going on inside the model, whereas explanations of model behaviour aim to explain what led to a specific behaviour or output by referencing essential attributes or influencers on that behaviour. It is not strictly necessary to understand the full set of relationships, dependencies, and weights of features within the model to explain model behaviour.
- Second, interpretability methods can be distinguished in how they conceptualise 'explanation'. Many methods conceptualise explanations as approximation models, which are a type of simpler, human interpretable model that is created to reliably approximate the functionality of a more complex 'black box' model. The approximation model itself is often and confusingly referred to as an explanation of the 'black box' model. This approach contrasts with the treatment of 'explanation' in philosophy of science and epistemology in which the term typically refers to explanatory statements that explain the causes of a given phenomenon.[211]

The usage of 'explanation' in this fashion can be confusing. Approximation models are best thought of as tools from which explanatory statements about the original model can be derived.[212] Explanatory statements themselves can be textual, quantitative, or visual, and report on several aspects of the model and its behaviours.

Further distinctions help classify different types of explanations and interpretability methods. A basic distinction in interpretability can be drawn between global and local interpretability. This distinction refers to the scope of the model or outputs a given interpretability or explanatory method aims to make human comprehensible. Global methods aim to explain the functionality of a model as a whole or across a particular set of outputs in terms of the significance of features, their dependencies or interactions, and their effect on outputs. In contrast, local methods can address, for example, the influence of specific areas of the input space or specific variables on one or more specific outputs of the model.

Models can be globally interpretable at a holistic or modular level.[213] Holistic global interpretability refers to models which are comprehensible to a human observer in the sense that the observer can follow the entire logic or functional steps taken by the model which lead to all possible outcomes of the model.[214] It should be possible for a

[211] Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY - FAT* '19 279–288 (2019).
[212] *Id.*
[213] MOLNAR, *supra* note 208.
[214] Guidotti et al., *supra* note 206.

single person to comprehend holistically interpretable models in their entirety.[215] An observer would have 'a holistic view of its features and each of the learned components such as weights, other parameters, and structures'.[216]

Given the limitations of human comprehension and short-term memory, global holistic interpretability is currently only practically achievable on relatively simple models with few features, interactions, or rules, or strong linearity and monotonicity.[217] For more complex models, global interpretability at a modular level may be feasible. This type of interpretability involves understanding a particular characteristic or segment of the model, for example the weights in a linear model, or the splits and leaf node predictions in a decision tree.[218]

With regards to local interpretability, a single output can be considered interpretable if the steps that led to it can be explained. Local interpretability does not strictly require that the entire series of steps be explained; rather, it can be sufficient to explain one or more aspects of the model that led to the output, such as a critically influential feature value.[219] A group of outputs is considered locally interpretable if the same methods to produce explanations of individual outputs can be applied to the group. Groups can also be explained by methods that produce global interpretability at a modular level.[220]

These distinctions lead to some initial conclusions about how AI can best explain itself to doctors and patients. At the point of adoption global explanations of model functionality seem appropriate to ensure a reliable fit between the intended use of the AI system in a given healthcare context, and the actual performance of the system. For explaining specific outputs or recommendations to patients, explanations of model behaviour formed as explanatory statements appear to strike the best fit between explaining the decision-making logic of the system while remaining comprehensible to expert and non-expert users alike. In this context methods such as 'counterfactual explanations' may be preferable as they facilitate debugging and testing of system performance by expert users while remaining comprehensible on an individual explanation level to non-expert patients.[221] To summarise, to make AI systems intelligible to patients, simple, local, contrastive explanations are preferable to global approximation explanations which can be difficult to understand and interpret.

An alternative but complementary approach is to use only intrinsically interpretable models in clinical care to enable health professionals to holistically understand systems and better explain them to their patients.[222] Implementing this approach would, however, create additional requirements for technical expertise in computer

[215] Lipton, *supra* note 164.
[216] MOLNAR, *supra* note 208 at 27.
[217] Guidotti et al., *supra* note 206.
[218] MOLNAR, *supra* note 208.
[219] *Id.*; Wachter, Mittelstadt, and Russell, *supra* note 172.
[220] MOLNAR, *supra* note 208.
[221] Wachter, Mittelstadt, and Russell, *supra* note 172.
[222] Cynthia Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, 1 NAT MACH INTELL 206–215 (2019).

science, statistics, and machine learning among health professionals which could be very difficult and perhaps unreasonable to meet in practice.

Public register of medical AI systems for transparency

As regards the issue of disclosure to patients of the usage of AI systems for operational and clinical purposes discussed in the section entitled "Transparency to health professionals and patients", the Parliamentary Assembly of the Council of Europe has recognised the importance of raising population awareness of uses of AI in healthcare to build trust with patients and ensure informed consent is possible in AI-mediated care. Specifically, their October 2020 report suggests that transparency of AI systems in healthcare "may require the establishment of a national health-data governance framework which could build on proposals from the international institutions. The latter include the Recommendation "Unboxing Artificial Intelligence: 10 steps to protect Human Rights" by the Council of Europe Commissioner for Human Rights (May 2019), the Ethics Guidelines for Trustworthy AI put forward by the European Union (April 2019), the OECD Recommendation and Principles on AI (May 2019) and the G20 Principles on Human-centred Artificial Intelligence (June 2019)."[223]

> Following these proposals and recommendations, a public database is seen as a key element to improve "algorithmic literacy" among the general public which is a fundamental precursor for exercising many human and legal rights.[224]

Insofar as the proposed framework is designed to increase population awareness of AI systems in healthcare, it can best be thought of as a type of public register for AI systems in healthcare. Registries are public lists of systems currently in use containing a standardised description of each system. Information included on registries varies but can include things like the intended usage or purpose of the system; its manufacturer or supplier; the underlying method(s) (e.g., deep learning, regression); any testing undergone both in terms of accuracy but also biases and other ethical and legal dimensions; a description of training and testing datasets; and an explanation of how predictions or outputs of the system are utilized by human decision-makers or otherwise integrated in existing services and decision-making processes.[225] Registries

[223] COUNCIL OF EUROPE, *supra* note 2.
[224] *Id.*
[225] Corinne Cath & Fieke Jansen, *Dutch Comfort: The limits of AI governance through municipal registers*, ARXIV PREPRINT ARXIV:2109.02944 (2021); Luciano Floridi, *Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki*, 33 PHILOSOPHY & TECHNOLOGY 541–546 (2020); Timnit Gebru et al., *Datasheets for Datasets* (2018), https://arxiv.org/abs/1803.09010 (last visited Oct 1, 2018); Margaret Mitchell et al., *Model Cards for Model Reporting*, PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY - FAT* '19 220–229 (2019); Sarah Holland et al., *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*, ARXIV:1805.03677 [CS] (2018), http://arxiv.org/abs/1805.03677 (last visited Oct 1, 2018).

also often have a feedback function to allow citizens to provide input on current and proposed uses of AI by public bodies and services.[226]

There are several examples of existing registries from municipal, national, and international public bodies. In 2020, Amsterdam and Helsinki launched public registries for AI and algorithmic systems used to deliver municipal services.[227] In November 2021, the UK Cabinet Office's Central Digital and Data Office launched a national algorithmic transparency standard which will effectively function as a type of public register.[228] Internationally, the recently proposed Artificial Intelligence Act contains a provision to create a public EU-wide database in which standalone high-risk AI applications must be registered.[229] The Council of Europe has an opportunity to complement these emerging transparency standards by introducing a public AI register for medical AI in member states which is aimed at patients to raise awareness of AI systems currently in use by their public health services.

Collection of sensitive data for bias and fairness auditing

Biases in AI systems linked to gaps in training and testing data could foreseeably motivate greater collection of sensitive data about legally protected groups for purposes of bias and fairness testing. It is a generally accepted fact, that in order to prevent discriminatory or biased outcomes, data on sensitive groups must be collected. Failure to collect this data will not prevent discrimination against protected groups, but arguably make it more difficult to detect.[230] Sensitive data is needed to test whether automated decision-making discriminated against groups based on protected attributes (e.g., data on race, disability, sexual orientation).[231] On the other hand, collecting such data has significant privacy implications. This is a legitimate concern and closely related to troubling historical experiences that significantly harmed specific groups in society.[232] For example, data collected for research and public purposes

[226] Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI, , VENTUREBEAT (2020), https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/ (last visited Dec 1, 2021).
[227] *Id.*
[228] UK government publishes pioneering standard for algorithmic transparency, , GOV.UK , https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency (last visited Dec 1, 2021).
[229] EUROPEAN COMMISSION, *supra* note 16 at Art. 51 and 60.
[230] SANDRA WACHTER, BRENT MITTELSTADT & CHRIS RUSSELL, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI* 34–35 (2020), https://papers.ssrn.com/abstract=3547922 (last visited Apr 19, 2020); Cynthia Dwork & Deirdre K. Mulligan, *It's not privacy, and it's not fair*, 66 STAN. L. REV. ONLINE 35 (2013); Cynthia Dwork et al., *Fairness Through Awareness*, ARXIV:1104.3913 [CS] (2011), http://arxiv.org/abs/1104.3913 (last visited Feb 15, 2016); Anupam Datta et al., *Proxy Non-Discrimination in Data-Driven Systems*, ARXIV:1707.08120 [CS] (2017), http://arxiv.org/abs/1707.08120 (last visited Jan 9, 2021); Kusner et al., *supra* note 189.
[231] Kusner et al., *supra* note 189; Chris Russell et al., *When worlds collide: integrating different counterfactual assumptions in fairness*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6396–6405 (2017).
[232] MAYER-SCHÖNBERGER AND CUKIER, *supra* note 31; For a US and EU comparison see Joris Van Hoboken, *From collection to use in privacy regulation? A forward looking comparison of European and US frameworks for personal data processing*, 231 EXPLORING THE BOUNDARIES OF BIG DATA (2016); For an international view 63 LEE A. BYGRAVE, DATA PRIVACY LAW: AN INTERNATIONAL PERSPECTIVE (2014); For

have contributed to eugenics in Europe, the UK[233] and the US,[234] genocide during WWII, racist immigration practices and the denial of basic human rights in the US,[235] justification of slavery,[236] forced sterilisation in the UK,[237] US, Germany and Puerto Rico from the early to the mid-20th Century,[238] punishment, castration and imprisonment of LGBT members,[239] and denial to women of equal rights and protection (e.g. sexual violence).[240] Clearly, privacy interests must be taken seriously when considering collection of sensitive personal data for purposes of bias testing.[241]

Setting these concerns aside for a moment, one could be tempted to think that the bias problems will naturally be solved by collecting more (sensitive) data and closing gaps in representation in training and testing datasets. However, fair and equal outcomes will not automatically result when representation gaps and other data biases are closed. Awareness of inequalities is not the same as rectifying them.[242] Rather, the persistence of social biases across Western societies suggest that significant political, social, and legal effort is needed to overcome them, rather than simply more data collection and testing.

Countering inequalities requires intentional and often cost intensive changes to decision processes, business models, and policies. To justify further collection and usage of sensitive data, it is necessary to first demonstrate serious commitment and political will to rectifying inequality. From a standard setting perspective, these observations suggest that any proposed collection of sensitive category data for the sake of testing medical AI systems form biases must have clear purpose limitations and confidentiality guarantees in place alongside a commitment to rectify social inequalities underlying biases discovered through testing. Operationalizing these commitments is not straightforward. The EU Artificial Intelligence Act, for example, proposes the creation of "regulatory sandboxes" in which AI providers can test their

---

an European view Sandra Wachter, *Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR*, 34 COMPUTER LAW & SECURITY REVIEW 436–449 (2018); Sandra Wachter, *The GDPR and the Internet of Things: a three-step transparency model*, 10 LAW, INNOVATION AND TECHNOLOGY 266–294 (2018); for a EU and German view see Mario Martini, Wiebke Fröhlich & Saskia Fritzsche, *Algorithmen als Herausforderung fu r die Rechtsordnung* (2017); for empirical evidence of mobile data collection see Reuben Binns et al., *Third party tracking in the mobile ecosystem*, *in* PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE 23–31 (2018); on online harms see Woods Lorna & Perrin William, *An updated proposal by Professor Lorna Woods and William Perrin*, https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/01/29121025/Internet-Harm-Reduction-final.pdf (last visited May 11, 2019).

[233] This happened until the 1930's, see RENI EDDO-LODGE, WHY I'M NO LONGER TALKING TO WHITE PEOPLE ABOUT RACE 20–21 (2020).

[234] JEAN HALLEY, AMY ESHLEMAN & RAMYA MAHADEVAN VIJAYA, SEEING WHITE: AN INTRODUCTION TO WHITE PRIVILEGE AND RACE 36 (2011).

[235] *Id.* at 25.

[236] *Id.* at 36–37.

[237] EDDO-LODGE, *supra* note 232 at 20–21.

[238] HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 233 at 36–38.

[239] JEAN HALLEY & AMY ESHLEMAN, SEEING STRAIGHT: AN INTRODUCTION TO GENDER AND SEXUAL PRIVILEGE 15–17 (2016).

[240] SAINI, *supra* note 178 at 233–235.

[241] For surveillance and chilling effects, see JON PENNEY, *Chilling Effects: Online Surveillance and Wikipedia Use* (2016), https://papers.ssrn.com/abstract=2769645 (last visited Dec 27, 2017).

[242] EDDO-LODGE, *supra* note 232 at 208.

systems for bias using special category data collected explicitly for testing purposes.[243] This proposal lacks the essential element of a commitment to rectify discovered inequalities.

[243] EUROPEAN COMMISSION, *supra* note 16 at Art. 53.

8 CONCLUDING REMARKS

Medical care is increasingly diffused across a variety of institutions, personnel, and technologies. The doctor-patient relationship has always adapted over time to advances in medicine, biomedical research, and care practices. At the same time, the capacity of AI to replace or augment human clinical expertise utilising highly complex analytics and unprecedented volumes and varieties of data suggests the impact of the technology on the doctor-patient relationship may be unprecedented.

The adoption of AI need not be a fundamental barrier to good doctor-patient relationships. AI has the potential to alter care relationships and displace responsibilities traditionally fulfilled by medical professionals, but this is not a foregone conclusion. The degree to which AI systems inhibit 'good' medical practice hinges upon the model of service. If AI is used solely to complement the expertise of health professionals bound by the fiduciary obligations of the doctor-patient relationship, the impact of AI on the trustworthiness and human quality of clinical encounters may prove to be minimal.

At the same time, if AI is used to heavily augment or replace human clinical expertise, its impact on the caring relationship is more difficult to predict. It is entirely possible that new, broadly accepted norms 'good' care will emerge through greater reliance on AI systems, with clinicians spending more time face-to-face with patients and relying heavily on automated recommendations.

The impact of AI on the doctor-patient relationship remains highly uncertain. We are unlikely to see a radical reconfiguration of care in the next five years in the sense of human expertise being replaced by artificial intelligence. With that said, developments like the COVID-19 pandemic and the increased pressures it has placed on health services may transform the mode of delivery of care if not the expertise behind it. Remote delivery of care, for example, may become increasingly commonplace even if diagnosis and treatment remain firmly in the hands of human health professionals.

A radical reconfiguration of the doctor-patient relationship of the type imagined by some commentators, in which artificial systems diagnose and treat patients directly with minimal interference from human clinicians, continues to seem far in the distance. Movement in this direction continues to hinge on proof of clinical efficacy which, as noted above, continues to prove a barrier to commercialisation and widespread adoption.[244] Likewise, new modes of clinical care would need to be derived that utilise the best aspects of human clinicians and artificial systems, implement appropriate safety and resilience checks, and minimise the weaknesses and implicit biases of both agents. Without due consideration of the implications of AI for medical practice, the "moral integrity of the doctor-patient relationship" may come to be dominated by institutional and external interests, with patient experiences of care suffering as a result.[245]

[244] Liu et al., *supra* note 112; Robbins and Brodwin, *supra* note 5.
[245] Bauer, *supra* note 136 at 90.

As AI is adopted across different healthcare systems and jurisdictions, it is important to remember that the moral obligations of the doctor-patient relationship are always affected and perhaps displaced by the introduction of new care providers. While technology continues to develop at a rapid pace, the patient's experience of illness (e.g., vulnerability, dependency) and expectations of the healing relationship do not radically or quickly change. The doctor-patient relationship is a keystone of 'good' medical practice, and yet it is seemingly being transformed into a doctor-patient-AI relationship. The challenge facing AI providers, regulators, and policymakers is to set robust standards and requirements for this new type of healing relationship to ensure patients' interests and the moral integrity of medicine as a profession are not fundamentally damaged by the introduction of disruptive emerging technologies.

APPENDIX: MEDICAL VIRTUES

Virtues are defined against the ends of the practice which they are meant to serve. For medicine, these ends are providing adequate care for a society, consisting of individual patients, in terms of physical and mental health and well-being. These ends are realised through the healing relationship, the nature of which introduces certain moral obligations.

As with all practices, phronesis or prudence is a central virtue in medicine, without which other virtues cannot be incorporated into behaviour through virtuous acts.[246] Justice, truthfulness and courage are also necessary to protect medicine from the corrupting power of medical institutions, including hospitals, paying organisations and government departments.[247] These three core virtues are necessary for continuous revision of standards of excellence and internal goods by practitioners, which requires critical self-reflection on the relationship between one's actions and the norms of the practice, or the institutional influence on the definition and realisation of norms.[248]

Justice is defined broadly as "the strict habit of rendering what is due to others,"[249] or "the virtue of rewarding desert and of repairing failures in rewarding desert within an already constituted community."[250] To be just, standards for treating people in a community must be "uniform and impersonal," meaning it is unjust to favour personal acquaintances. In social or national healthcare systems, justice can be applied to the distribution of medical resources (e.g., pharmaceuticals, treatments, clinical encounters) in a manner fair to all stakeholders. Justice is not merely a quantitative notion, by which all stakeholders receive an equal share, but instead requires matching resources to the needs of the patient and making judgments between the relative importance of different needs.

Fidelity to trust and beneficence can also be understood as core virtues unique to medicine because of the need for trust in healing relationships.[251] A trusting relationship needs to develop over time between the virtuous doctor and patient, in which the values, expectations and thoughts on illness and appropriate medical care are shared. The patient must at a minimum believe the doctor is acting beneficently, or in his interests and well-being, to some degree for trust to exist.[252]

---

[246] MACINTYRE, *supra* note 122 at 154; G. Widdershoven & Lieke Van der Scheer, *Theory and methodology of empirical ethics: a pragmatic hermeneutic perspective*, *in* EMPIRICAL ETHICS IN PSYCHIATRY 23–36 (2008), http://books.google.co.uk/books?hl=en&lr=&id=Lvq0lkDyEBQC&oi=fnd&pg=PA23&dq=Theory+and+methodology+of+empirical+ethics:+a+pragmatic+hermeneutic+perspective&ots=IXt3OC6Obh&sig=EU-idi92-6EzBI6uTp8UNReq4AY#v=onepage&q&f=false; PELLEGRINO AND THOMASMA, *supra* note 120.
[247] MACINTYRE, *supra* note 122 at 192.
[248] *Id.* at 191.
[249] PELLEGRINO AND THOMASMA, *supra* note 120 at 92.
[250] MACINTYRE, *supra* note 122 at 156.
[251] PELLEGRINO AND THOMASMA, *supra* note 120 at 71, 156.
[252] *Id.* at 156.

Other virtues include compassion, fortitude, integrity and temperance. Compassion is the trait of a doctor which allows him to 'enter the perspective' of the patient, to understand how the patient's values, expectations of care, social, emotional and physical well-being affect his experience of illness, and to customise his care and recommendations to the needs of each patient as a unique individual.[253] Compassion may also necessitate the promotion of health-related values and deliberation with the patient to convince him of the best intervention in terms of fit between health outcomes as perceived by the doctor and the patient's values.[254]

Fortitude is a form of moral courage, by which an individual is willing to "suffer personal harm for the sake of a moral good" such as a doctor refusing to act in accordance with institutional rules which would be detrimental to his patient's well-being, risking harm to his career and professional membership.[255] Fortitude can create an obligation for doctors to speak out against the potential harms of new institutional policies, technologies or treatments for their patients. Temperance is the restriction of behaviour in a practice to meet the moral obligations of that practice. It can be used synonymously with virtue itself but is distinct as a character trait of the virtuous doctor who suppresses self-interest in treating patients. Without such restraint other virtues cannot be practiced.[256]

Integrity is the possession of all virtues combined with the ability to discern between moral principles in choosing appropriate actions conducive to the good of medicine in different situations.[257] It is the core virtue of the narrative quest for the good life, and can be seen in a life of virtuous behaviour.[258] Integrity can be exercised when a doctor promotes the patient's interests and welfare in the face of institutional pressure, for example by not sending a patient home early from hospital.[259] Edgar and Pattison define integrity as "the capacity to deliberate and reflect usefully in the light of context, knowledge, experience and information (that of self and other) on complex and conflicting factors bearing on action or potential action."[260] Integrity is therefore perhaps indistinguishable from phronesis, temperance and fortitude.

[253] *Id.* at 79, 81.
[254] Emanuel and Emanuel, *supra* note 123 at 2226.
[255] PELLEGRINO AND THOMASMA, *supra* note 120 at 109.
[256] *Id.* at 117.
[257] *Id.* at 127.; Edgar and Pattison, *supra* note 145 at 102.
[258] MACINTYRE, *supra* note 122.
[259] Edgar and Pattison, *supra* note 145 at 94.
[260] *Id.* at 102.

BRIEFING

EU Legislation in Progress

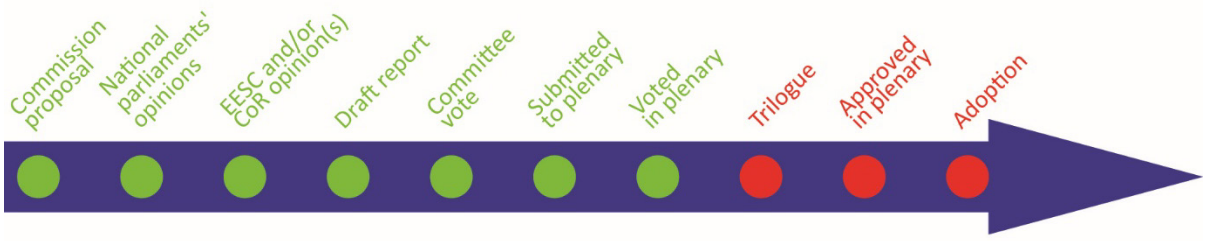
European Parliament

Artificial intelligence act

OVERVIEW

The European Commission tabled a proposal for an EU regulatory framework on artificial intelligence (AI) in April 2021. The draft AI act is the first ever attempt to enact a horizontal regulation for AI. The proposed legal framework focuses on the specific utilisation of AI systems and associated risks. The Commission proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored on a 'risk-based approach'. Some AI systems presenting 'unacceptable' risks would be prohibited. A wide range of 'high-risk' AI systems would be authorised, but subject to a set of requirements and obligations to gain access to the EU market. Those AI systems presenting only 'limited risk' would be subject to very light transparency obligations. The Council agreed the EU Member States' general position in December 2021. Parliament voted on its position in June 2023. EU lawmakers are now starting negotiations to finalise the new legislation, with substantial amendments to the Commission's proposal including revising the definition of AI systems, broadening the list of prohibited AI systems, and imposing obligations on general purpose AI and generative AI models such as ChatGPT.

Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts		
Committees responsible:	Internal Market and Consumer Protection (IMCO) and Civil Liberties, Justice and Home Affairs (LIBE) (jointly under Rule 58)	COM(2021)206 21.4.2021 2021/0106(COD)
Rapporteurs:	Brando Benifei (S&D, Italy) and Dragoş Tudorache (Renew, Romania)	
Shadow rapporteurs:	Deirdre Clune, Axel Voss (EPP); Petar Vitanov (S&D); Svenja Hahn, (Renew); Sergey Lagodinsky, Kim Van Sparrentak (Greens/EFA); Rob Rooken, Kosma Złotowski (ECR); Jean-Lin Lacapelle, Jaak Madison (ID); Cornelia Ernst, Kateřina Konecna (The Left)	Ordinary legislative procedure (COD) (Parliament and Council on equal footing – formerly 'co-decision')
Next steps expected:	Trilogue negotiations	



EPRS | European Parliamentary Research Service

Author: Tambiama Madiega; Graphic: Samy Chahri
Members' Research Service

PE 698.792 – June 2023

EN

Introduction

AI technologies are expected to bring a wide array of **economic and societal benefits** to a wide range of sectors, including environment and health, the public sector, finance, mobility, home affairs and agriculture. They are particularly useful for improving prediction, for optimising operations and resource allocation, and for personalising services.[1] However, the implications of AI systems for **fundamental rights** protected under the EU Charter of Fundamental Rights, as well as the **safety risks** for users when AI technologies are embedded in products and services, are raising concern. Most notably, AI systems may jeopardise fundamental rights such as the right to non-discrimination, freedom of expression, human dignity, personal data protection and privacy.[2]

Given the fast development of these technologies, in recent years AI regulation has become a central policy question in the European Union (EU). Policy-makers pledged to develop a **'human-centric' approach to AI** to ensure that Europeans can benefit from new technologies developed and functioning according to the EU's values and principles.[3] In its 2020 White Paper on Artificial Intelligence, the European Commission committed to **promote the uptake of AI** and **address the risks associated** with certain uses of this new technology. While the European Commission initially adopted a **soft-law approach**, with the publication of its non-binding 2019 Ethics Guidelines for Trustworthy AI and Policy and investment recommendations, it has since shifted towards a **legislative approach**, calling for the adoption of harmonised rules for the development, placing on the market and use of AI systems.[4]

AI regulatory approach in the world. While the United States of America (USA) had initially taken a lenient approach towards AI, calls for regulation have recently been mounting. The Cyberspace Administration of China is also consulting on a proposal to regulate AI, while the UK is working on a set of pro-innovation regulatory principles. At international level, the Organisation for Economic Co-operation and Development (OECD) adopted a (non-binding) Recommendation on AI in 2019, UNESCO adopted Recommendations on the Ethics of AI in 2021, and the Council of Europe is currently working on an international convention on AI. Furthermore, in the context of the newly established EU-US tech partnership (the Trade and Technology Council), the EU and USA are seeking to develop a mutual understanding on the principles underlining trustworthy and responsible AI. EU lawmakers issued a joint statement in May 2023 urging President Biden and European Commission President Ursula von der Leyen to convene a summit to find ways to control the development of advanced AI systems such as ChatGPT.

Parliament's starting position

Leading the EU-level debate, the European Parliament called on the European Commission to assess the impact of AI and to draft an EU framework for AI, in its wide-ranging 2017 recommendations on civil law rules on robotics. More recently, in 2020 and 2021, the Parliament adopted a number of non-legislative resolutions calling for EU action, as well as two legislative resolutions calling for the adoption of EU legislation in the field of AI. A first legislative resolution asked that the Commission establish a legal framework of ethical principles for the development, deployment and use of AI, robotics and related technologies in the Union. A second legislative resolution called for harmonisation of the legal framework for civil liability claims and imposition of a regime of strict liability on operators of high-risk AI systems. Furthermore, the Parliament adopted a series of recommendations calling for a common EU approach to AI in the intellectual property, criminal law, education, culture and audiovisual areas, and regarding civil and military AI uses.

Council starting position

In the past, the Council has repeatedly called for the adoption of common AI rules, including in 2017 and 2019. More recently, in 2020, the Council called upon the Commission to put forward concrete proposals that take existing legislation into account and follow a risk-based, proportionate and, if necessary, regulatory approach. Furthermore, the Council invited the EU and the Member States to

consider effective measures for identifying, predicting and responding to the potential impacts of digital technologies, including AI, on fundamental rights.

Preparation of the proposal

Following the White Paper on Artificial Intelligence[5] adopted in February 2020, the Commission launched a broad public consultation in 2020 and published an Impact Assessment of the regulation on artificial intelligence, a supporting study and a draft proposal, which received feedback from a variety of stakeholders.[6] In its impact assessment, the Commission identifies **several problems** raised by the development and use of AI systems, due to their specific characteristics.[7]

The changes the proposal would bring

The draft AI act has been designed as a **horizontal EU legislative instrument** applicable to all AI systems placed on the market or used in the Union.

Purpose, legal basis and scope

The **general objective** of the proposed AI act unveiled in April 2021 is to ensure the proper functioning of the single market by creating the conditions for the development and use of trustworthy AI systems in the Union. The draft sets out a harmonised legal framework for the development, placing on the Union market, and the use of AI products and services. In addition, the AI act proposal seeks to achieve a set of **specific objectives**: (i) ensure that AI systems placed on the EU market are safe and respect existing EU law, (ii) ensure legal certainty to facilitate investment and innovation in AI, (iii) enhance governance and effective enforcement of EU law on fundamental rights and safety requirements applicable to AI systems, and (iv) facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.[8]

The new AI framework, based on Article 114[9] and Article 16[10] of the Treaty on the Functioning of the European Union (TFEU), would enshrine a **technology-neutral definition of AI systems** and adopt a **risk-based approach**, which lays down different **requirements and obligations** for the development, placing on the market and use of AI systems in the EU. In practice, the proposal defines common mandatory requirements applicable to the design and development of AI systems before they are placed on the market and harmonises the way ex-post controls are conducted. The proposed AI act would complement existing and forthcoming, horizontal and sectoral EU safety regulation.[11] The Commission proposes to follow the logic of the new legislative framework (NLF), i.e. the EU approach to ensuring a range of products comply with the applicable legislation when they are placed on the EU market through conformity assessments and the use of CE marking.

The new rules would apply primarily to **providers of AI systems established within the EU or in a third country** placing AI systems on the EU market or putting them into service in the EU, as well as to **users of AI systems located in the EU**.[12] To prevent circumvention of the regulation, the new rules would also apply to **providers and users of AI systems located in a third country** where the output produced by those systems is used in the EU.[13] However, the draft regulation does not apply to AI systems developed or used exclusively for military purposes, to public authorities in a third country, nor to international organisations, or authorities using AI systems in the framework of international agreements for law enforcement and judicial cooperation.

Definitions

No single definition of artificial intelligence is accepted by the scientific community and the term 'AI' is often used as a 'blanket term' for various computer applications based on different techniques, which exhibit capabilities commonly and currently associated with human intelligence.[14] The High Level Expert Group on AI proposed a baseline definition of AI that is increasingly used in the scientific literature, and the Joint Research Centre has established an operational definition of AI based on a taxonomy that maps all the AI subdomains from a political, research and industrial

perspective. However, the Commission found that the **notion of an AI system** should be more clearly defined, given that the determination of what an 'AI system' constitutes is crucial for the allocation of legal responsibilities under the new AI framework. The Commission therefore proposes to establish a legal definition of 'AI system' in EU law, which is largely based on a definition already used by the OECD.[15] Article 3(1) of the draft act states that '**artificial intelligence system**' means:

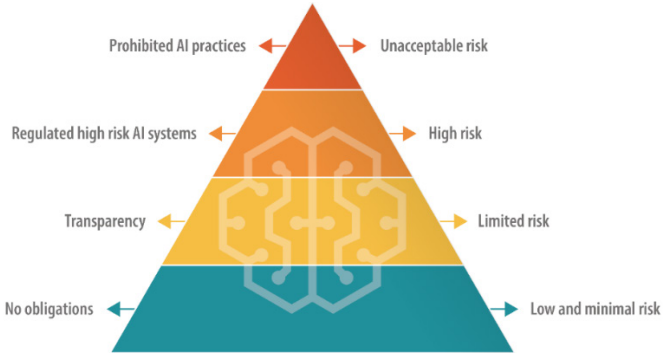
> *...software that is developed with [specific] techniques and approaches [listed in Annex 1] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.*[16]

Annex 1 of the proposal lays out a **list of techniques and approaches** that are used today to develop AI. Accordingly, the notion of 'AI system' would refer to a range of software-based technologies that encompasses '**machine learning**', '**logic and knowledge-based**' systems, and '**statistical**' approaches. This broad definition covers AI systems that can be used on a stand-alone basis or as a component of a product. Furthermore, the proposed legislation aims to be future-proof and cover current and future AI technological developments. To that end, the Commission would complement the Annex 1 list with new approaches and techniques used to develop AI systems as they emerge – through the adoption of **delegated acts** (Article 4).

Furthermore, Article 3 provides a long **list of definitions** including that of 'provider' and 'user' of AI systems (covering both public and private entities), as well as 'importer' and 'distributor', 'emotion recognition', and 'biometric categorisation'.

Risk-based approach

Pyramid of risks



Data source: European Commission.

The use of AI, with its specific characteristics (e.g. opacity, complexity, dependency on data, autonomous behaviour), can adversely affect a number of fundamental rights and users' safety. To address those concerns, the draft AI act follows a **risk-based approach** whereby legal intervention is tailored to concrete level of risk. To that end, the draft AI act distinguishes between AI systems posing (i) **unacceptable risk**, (ii) **high risk**, (iii) **limited risk**, and (iv) **low or minimal risk**. AI applications would be regulated only as strictly necessary to address specific levels of risk.[17]

Unacceptable risk: Prohibited AI practices

Title II (Article 5) of the proposed AI act explicitly **bans harmful AI practices** that are considered to be a clear threat to people's safety, livelihoods and rights, because of the 'unacceptable risk' they create. Accordingly, it would be prohibited to place on the market, put into services or use in the EU:

- AI systems that deploy harmful manipulative 'subliminal techniques';
- AI systems that exploit specific vulnerable groups (physical or mental disability);
- AI systems used by public authorities, or on their behalf, for social scoring purposes;
- 'Real-time' remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.[18]

High risk: Regulated high-risk AI systems

Title III (Article 6) of the proposed AI act regulates 'high-risk' AI systems that create adverse impact on people's safety or their fundamental rights. The draft text distinguishes between two categories of high-risk AI systems.

- Systems used as a safety component of a product or falling under EU health and safety harmonisation legislation (e.g. toys, aviation, cars, medical devices, lifts).
- Systems deployed in **eight specific areas** identified in Annex III, which the Commission could update as necessary through **delegated acts** (Article 7):
  - Biometric identification and categorisation of natural persons;
  - Management and operation of critical infrastructure;
  - Education and vocational training;
  - Employment, worker management and access to self-employment;
  - Access to and enjoyment of essential private services and public services and benefits;
  - Law enforcement;
  - Migration, asylum and border control management;
  - Administration of justice and democratic processes.

All of these high-risk AI systems would be subject to a set of new rules including:

Requirement for an ex-ante conformity assessment: Providers of high-risk AI systems would be required to register their systems in an **EU-wide database** managed by the Commission before placing them on the market or putting them into service. Any AI products and services governed by existing product safety legislation will fall under the existing third-party conformity frameworks that already apply (e.g. for medical devices). Providers of AI systems not currently governed by EU legislation would have to conduct their own conformity assessment (**self-assessment**) showing that they comply with the new requirements and can use **CE marking**. Only high-risk AI systems used for biometric identification would require a conformity assessment by a 'notified body'.

Other requirements: Such high-risk AI systems would have to comply with a range of requirements particularly on risk management, testing, technical robustness, data training and data governance, transparency, human oversight, and cybersecurity (Articles 8 to 15). In this regard, providers, importers, distributors and users of high-risk AI systems would have to fulfil a range of obligations. Providers from outside the EU will require an **authorised representative** in the EU to (inter alia), ensure the conformity assessment, establish a post-market monitoring system and take corrective action as needed. AI systems that conform to the new **harmonised EU standards**, currently under development, would benefit from a presumption of conformity with the draft AI act requirements.[19]

Facial recognition: AI powers the use of biometric technologies, including facial recognition technologies (FRTs), which are used by private or public actors for verification, identification and categorisation purposes. In addition to the existing applicable legislation (e.g. data protection and non-discrimination), the draft AI act proposes to introduce new rules for FRTs and differentiate them according to their 'high-risk' or 'low-risk' usage characteristics. The use of real-time facial recognition systems in publicly accessible spaces for the purpose of law enforcement would be prohibited, unless Member States choose to authorise them for important public security reasons, and the appropriate judicial or administrative authorisations are granted. A wide range of FRTs used for purposes other than law enforcement (e.g. border control, market places, public transport and even schools) could be permitted, subject to a conformity assessment and compliance with safety requirements before entering the EU market.[20]

Limited risk: Transparency obligations

AI systems presenting 'limited risk', such as **systems that interacts with humans** (i.e. chatbots), **emotion recognition systems**, **biometric categorisation systems**, and AI systems that generate or manipulate image, audio or video content (i.e. **deepfakes**), would be subject to a limited set of transparency obligations.

Low or minimal risk: No obligations

All other AI systems presenting only low or minimal risk could be developed and used in the EU without conforming to any additional legal obligations. However, the proposed AI act envisages the creation of **codes of conduct** to encourage providers of non-high-risk AI systems to voluntarily apply the mandatory requirements for high-risk AI systems.

Governance, enforcement and sanctions

The proposal requires Member States to designate one or more competent authorities, including a **national supervisory authority**, which would be tasked with supervising the application and implementation of the regulation, and establishes a **European Artificial Intelligence Board** (composed of representatives from the Member States and the Commission) at EU level. National **market surveillance authorities** would be responsible for assessing operators' compliance with the obligations and requirements for high-risk AI systems. They would have access to confidential information (including the source code of the AI systems) and subject to binding confidentiality obligations. Furthermore, they would be required to take any **corrective measures** to prohibit, restrict, withdraw or recall AI systems that do not comply with the AI act, or that, although compliant, present a risk to health or safety of persons or to fundamental rights or other public interest protection. In case of persistent non-compliance, Member States will have to take all appropriate measures to restrict, prohibit, recall or withdraw the high-risk AI system at stake from the market.

Administrative **fines** of varying scales (up to €30 million or 6 % of the total worldwide annual turnover), depending on the severity of the infringement, are set as sanctions for non-compliance with the AI act. Member States would need to lay down rules on penalties, including administrative fines and take all measures necessary to ensure that they are properly and effectively enforced.

Measures to support innovation

The Commission proposes that Member States, or the European Data Protection Supervisor, could establish a **regulatory sandbox**, i.e. a controlled environment that facilitates the development, testing and validation of innovative AI systems (for a limited period of time) before they are put on the market. Sandboxing will enable participants to use personal data to foster AI innovation, without prejudice to the GDPR requirements. Other measures are tailored specifically to small-scale providers and **start-ups**

Advisory committees

The European Economic and Social Committee adopted its opinion on the proposed artificial intelligence act on 22 September 2021.

National parliaments

The deadline for the submission of reasoned opinions on the grounds of subsidiarity was 2 September 2021. Contributions were received from the Czech Chamber of Deputies and the Czech Senate, the Portuguese Parliament, the Polish Senate and the German Bundesrat.

Stakeholder views[21]

Definitions

Definitions are a contentious point of discussion among stakeholders. The Big Data Value Association, an industry-driven international not–for-profit organisation, stresses that the definition of AI systems is quite broad and would cover far more than what is subjectively understood as AI, including the simplest search, sorting and routing algorithms, which would consequently be subject to new rules. Furthermore, they ask for clarification of how components of larger AI systems (such

as pre-trained AI components from other manufacturers or components not released separately), should be treated. AmCham, the American Chamber of Commerce in the EU, suggests avoiding over-regulation by adopting a narrower definition of AI systems, focusing strictly on high-risk AI applications (and not extended to AI applications that are not high-risk, or software in general). AccessNow, an association defending users' digital rights argues the definitions of 'emotion recognition' and 'biometric categorisation' are technically flawed, and recommends adjustments.

Risk-based approach

While they generally welcome the proposed AI act's risk-based approach, some stakeholders support wider prohibition and regulation of AI systems. Civil rights organisations call for a ban on indiscriminate or arbitrarily targeted use of biometrics in public or publicly accessible spaces, and for restrictions on the uses of AI systems, including for border control and predictive policing. AccessNow argues that the provisions concerning prohibited AI practices (Article 5) are too vague, and proposes a wider ban on the use of AI to categorise people based on physiological, behavioural or biometric data, for emotion recognition, as well as dangerous uses in the context of policing, migration, asylum, and border management. Furthermore, they call for stronger impact assessment and transparency requirements.

The European Enterprises Alliance stresses that there is general uncertainty about the roles and responsibilities of the different actors in the AI value chain (developers, providers, and users of AI systems). This is particularly challenging for companies providing general purpose application programming interfaces or open-source AI models that are not specifically intended for high-risk AI systems but are nevertheless used by third parties in a manner that could be considered high-risk. They also call for 'high-risk' to be redefined, based on the measurable harm and potential impact. AlgorithmWatch underlines that the applicability of specific rules should not depend on the type of technology, but on the impact it has on individuals and society. They call for the new rules to be defined according to the impact of the AI systems and recommend that every operator should conduct an impact assessment that assesses the system's risk levels on a case-by-case basis. Climate Change AI calls for climate change mitigation and adaptation to be taken into account in the classification rules for high-risk AI systems and impose environmental protection requirements.

Consumer protection

The European Consumer Organisation, BEUC, stresses that the proposal requires substantial improvement to guarantee consumer protection. The organisation argues that the proposal should have a broader scope and impose basic principles and obligations (e.g. on fairness, accountability and transparency) upon all AI systems, as well as prohibiting more comprehensively harmful practices (such as private entities' use of social scoring and of remote biometric identification systems in public spaces). Furthermore, consumers should be granted a strong set of rights, effective remedies and redress mechanisms, including collective redress.

Impact on investments and SMEs

There are opposing views on the impact of the proposed regulation on investment. A study by the Centre for Data Innovation (representing large online platforms) highlights that the compliance costs incurred under the proposed AI act would likely provoke a chilling effect on investment in AI in Europe, and could particularly deter small and medium-sized enterprises (SMEs) from developing high-risk AI systems. According to the Centre for Data Innovation, the AI act would cost the European economy €31 billion over the next five years and reduce AI investments by almost 20 %. However, such estimates of the compliance costs are challenged by the experts from the Centre for European Policy Studies, as well as by other economists. The European Digital SME Alliance warns against overly stringent conformity requirements, asks for effective representation of SMEs in the standards-setting procedures and for making sandboxes mandatory in all EU Member States.

Academic and other views

While generally supporting the Commission's proposal, critics call for amendments, including revising the 'AI systems' definition, ensuring a better allocation of responsibility, strengthening enforcement mechanisms and fostering democratic participation.[22] Among the main issues are:

AI systems definition

The legal definition of 'AI systems' contained in the proposed AI act has been heavily criticised. Smuha and others warn the definition lacks clarity and may lead to legal uncertainty, especially for some systems that would not qualify as AI systems under the draft text, while their use may have an adverse impact on fundamental rights.[23] To address this issue, the authors propose to **broaden the scope of the legislation** to explicitly include all computational systems used in the identified high-risk domains, regardless of whether they are considered to be AI. According to the authors, the advantage would be in making application of the new rules more dependent on the domain in which the technology is used and the fundamental rights-related risks, rather than on a specific computational technique. Ebers and others consider that the scope of 'AI systems' is overly broad, which may lead to **legal uncertainty** for developers, operators, and users of AI systems and ultimately to over-regulation.[24] They call on EU law-makers to exempt AI systems developed and used for **research purposes** and **open-source software** (OSS) from regulation. Other commentators question whether the proposed definition of 'AI systems' is truly **technology neutral** as it refers primarily to 'software', omitting potential future AI developments.

Risk-based approach

Academics also call for amendments, warning that the risk-based approach proposed by the Commission would not ensure a high level of protection of fundamental rights. Smuha and others argue that the proposal does not always accurately recognise the wrongs and harms associated with different kinds of AI systems and therefore does not appropriately allocate responsibility. Among other things, they recommend adding a procedure that enables the Commission to **broaden the list of prohibited AI systems**, and propose banning existing manipulative AI systems (e.g. deepfakes), social scoring and some biometrics. Ebers and others call for a **more detailed classification of risks** to facilitate industry self-assessment and support, as well as **prohibiting more AI systems** (e.g. biometrics), including in the context of **private use**. Furthermore, some highlight that the draft legislation does not address **systemic sustainability risks** created by AI especially in the area of climate and environmental protection.[25]

Experts seem particularly concerned by the implementation of Article 5 (prohibited practices) and Article 6 (regulated high-risk practices). One of the major concerns raised is that the rules on prohibited and high-risk practices may prove ineffective in practice, because the risk assessment is left to provider **self-assessment**. Veale and Zuiderveen Borgesius warn that most providers can arbitrarily classify most high-risk systems as adhering to the rules using self-assessment procedures alone. Smuha and others recommend exploring whether certain high-risk systems would not benefit from a conformity assessment carried out by an **independent entity** prior to their deployment.

Biometrics regulation. A study commissioned by the European Parliament recommends, inter alia, to empower the Commission to adapt the list of prohibited AI practices periodically, under the supervision of the European Parliament, and the adoption of a more comprehensive list of 'restricted AI applications' (comprising real-time remote biometric identification without limitation for law enforcement purposes). Regulation of facial recognition technologies (FRTs) is one of the most contentious issues.[26] The European Data Protection Supervisor (EDPS) and the European Data Protection Board (EDPB) have called for a general ban on any uses of AI for the automated recognition of human features in publicly accessible spaces.

Governance structure and enforcement and redress mechanisms

Ebers and others stress that the AI act **lacks effective enforcement structures**, as the Commission proposes to leave the preliminary risk assessment, including the qualification as high-risk, to the providers' self-assessment. They also raise concerns about the excessive delegation of regulatory power to private European standardisation organisations (ESOs), due to the lack of democratic oversight, the impossibility for stakeholders (civil society organisations, consumer associations) to influence the development of standards, and the lack of judicial means to control them once they have been adopted. Instead, they recommend that the AI act codifies a set of legally binding requirements for high-risk AI systems (e.g. prohibited forms of algorithmic discrimination), which ESOs may specify through harmonised standards. Furthermore, they advocate that European policy-makers should **strengthen democratic oversight of the standardisation process**.

Commentators deplore a crucial gap in the AI act, which does not provide for **individual enforcement rights**. Ebers and others stress that individuals affected by AI systems and civil rights organisations have no **right to complain** to market surveillance authorities or to sue a provider or user for failure to comply with the requirements. Similarly, Veale and Zuiderveen Borgesius warn that, while some provisions of the draft legislation aim to impose obligations on AI systems users, there is **no mechanism for complaint or judicial redress** available to them. Smuha and others recommend amending the proposal to include, inter alia, an **explicit right of redress for individuals** and **rights of consultation and participation for EU citizens** regarding the decision to amend the list of high-risk systems in Annex III.

It has also been stressed that the text as it stands **lacks proper coordination** mechanisms between authorities, in particular concerning **cross-border infringement**. Consequently, the competence of the relevant authorities at national level should be clarified. Furthermore, guidance would be desirable on how to ensure compliance with transparency and information requirements, while simultaneously **protecting intellectual property rights and trade secrets** (e.g. to what extent the source code must be disclosed), not least to avoid diverging practices in the Member States.

Legislative process

The **Council** adopted its common position in December 2022. The Council's proposes, inter alia to:

- narrow the definition of AI systems to systems developed through machine learning approaches and logic- and knowledge-based approaches;
- extend to private actors the prohibition on using AI for social scoring, and add cases when the use of 'real-time' remote biometric identification systems in publicly accessible spaces could exceptionally be allowed;
- impose requirements on general purpose AI systems by means of implementing acts;
- add new provisions to take into account situations where AI systems can be used for many different purposes (general purpose AI); and
- simplify the compliance framework for the AI Act and strengthen, in particular, the role of the AI Board.

In **Parliament**, the file was assigned jointly (under Rule 58) to the Committee on Internal Market and Consumer Protection (IMCO) and the Committee on Civil Liberties, Justice and Home Affairs (LIBE), with Brando Benifei (S&D, Italy) and Dragos Tudorache, Renew, Romania) appointed as rapporteurs. In addition, the Legal Affairs Committee (JURI), the Committee on Industry, Research and Energy (ITRE) and the Committee on Culture and Education (CULT) are each associated to the legislative work under Rule 57, with shared and/or exclusive competences for specific aspects of the proposal. Parliament adopted its negotiating position (499 votes in favour, 28 against and 93 abstentions) on 14 June 2023, with substantial amendments to the Commission's text, including:

- **Definitions**. Parliament amended the definition of AI systems to align it with the definition agreed by the OECD. Furthermore, Parliament enshrines a definition of

'general purpose AI system' and 'foundation model' in EU law.

- **Prohibited practices**. Parliament substantially amended the list of AI systems prohibited in the EU. Parliament wants to ban the use of biometric identification systems in the EU for both real-time and ex-post use (except in cases of severe crime and pre-judicial authorisation for ex-post use) and not only for real-time use, as proposed by the Commission. Furthermore, Parliament wants to ban all biometric categorisation systems using sensitive characteristics (e.g. gender, race, ethnicity, citizenship status, religion, political orientation); predictive policing systems (based on profiling, location or past criminal behaviour); emotion recognition systems (used in law enforcement, border management, workplace, and educational institutions); and AI systems using indiscriminate scraping of biometric data from social media or CCTV footage to create facial recognition databases.
- **High-risk AI systems**. While the Commission proposed to automatically categorise as high-risk all systems in certain areas or use cases, Parliament adds the additional requirement that the systems must pose a 'significant risk' to qualify as high-risk. AI systems that risk harming people's health, safety, fundamental rights or the environment would be considered as falling within high-risk areas. In addition, AI systems used to influence voters in political campaigns and AI systems used in recommender systems displayed by social media platforms, designated as very large online platforms under the Digital Services Act, would be considered high-risk systems. Furthermore, Parliament imposes on those deploying a high-risk system in the EU an obligation to carry out a fundamental rights impact assessment.
- **General-purpose AI, generative AI and foundation models.** Parliament sets a layered regulation of general-purpose AI. Parliament imposes an obligation on providers of foundation models to ensure robust protection of fundamental rights, health, safety, the environment, democracy and the rule of law. They would be required to assess and mitigate the risks their models entail, comply with some design, information and environmental requirements and register such models in an EU database. Furthermore, generative foundation AI models (such as ChatGPT) that use large language models (LLMs) to generate art, music and other content would be subject to stringent transparency obligations. Providers of such models and of generative content would have to disclose that the content was generated by AI not by humans, train and design their models to prevent generation of illegal content and publish information on the use of training data protected under copyright law. Finally, all foundation models should provide all necessary information for downstream providers to be able to comply with their obligations under the AI act.
- **Governance and enforcement**. National authorities' competences would be strengthened, as Parliament gives them the power to request access to both the trained and training models of the AI systems, including foundation models. Parliament also proposes to establish an AI Office, a new EU body to support the harmonised application of the AI act, provide guidance and coordinate joint cross-border investigations. In addition, Members seek to strengthen citizens' rights to file complaints about AI systems and receive explanations of decisions based on high-risk AI systems that significantly impact their rights.
- **Research and innovation**. To support innovation, Parliament agrees that research activities and the development of free and open-source AI components would be largely exempted from compliance with the AI act rules.

Policy debate latest issues. The recent and rapid development of general-purpose artificial intelligence technologies has framed the policy debate around, inter alia, defining general-purpose AI models, the application of the EU copyright framework to **generative AI**, how to ensure foundation models' compliance with AI Act principles, and the design of efficient auditing procedures for **large language models** (LLMs). A risk of **over-regulation** detrimental for investment in AI in the EU has been identified should overly stringent obligations of risk assessment, mitigation and management be imposed on **foundation models** and on SMEs. How to set pro-competitive rules for sandboxing and open-source AI systems has also been discussed. While there are concerns that AI poses societal-scale risks similar to nuclear weapons, calls for a pause in AI development have been made by civil society organisations, AI experts and tech executives. The question how to address dual-use and military AI applications has also been raised. Furthermore, given EU regulation will take time to take effect, the adoption of voluntary codes of conduct and of an AI Pact are envisaged to mitigate the potential downsides of generative AI. A pressing issue is to set a **common** terminology so that lawmakers around the globe have the same understanding of the technologies they need to address.

EP SUPPORTING ANALYSIS

General-purpose artificial intelligence, EPRS, Madiega T., March 2023.

Biometric Recognition and Behavioural Detection, European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs, August 2021.

Regulating facial recognition in the EU, EPRS, Madiega T. A. and Mildebrath H. A., September 2021.

Artificial intelligence in criminal law, EPRS, Voronova S., September 2021.

Artificial Intelligence Act: Initial Appraisal of the European Commission Impact Assessment, Dalli H., EPRS, July 2021.

Artificial intelligence at EU borders: Overview of applications and key issues, Dumbrava C., EPRS, July 2021.

OTHER SOURCES

Artificial Intelligence Act, European Parliament, Legislative Observatory (OEIL).

Ebers M., and others, The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS), J 4, no 4: 589-603, October 2021.

Smuha N., and others, How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act, Elsevier, August 2021.

Veale M., Zuiderveen Borgesius F., Demystifying the draft EU AI Act, 22(4) *Computer Law Review International*, July 2021.

ENDNOTES

1 See European Commission, Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) 2021/0106 (COD), Explanatory memorandum (Commission proposal for an AI act). While the exact definition of AI is highly contested (see below), it is generally acknowledged that AI combines a range of technologies including machine-learning techniques, robotics and automated decision-making systems.

2 See for instance, High-Level Expert Group, Ethics Guidelines for Trustworthy AI, 2019.

3 See European Commission, Communication on Building Trust in Human-Centric Artificial Intelligence, COM(2019) 168.

4 See European Commission, Communication on Fostering a European approach to Artificial Intelligence, COM(2021) 205.

5 See European Commission, White Paper on Artificial Intelligence, COM(2020) 65 final.

6 For an overview see H. Dalli, Artificial intelligence act, Initial Appraisal of a European Commission Impact Assessment, EPRS, European Parliament, 2021.

7 According to the Commission impact assessment, the five specific characteristics of AI are (i) opacity (limited ability of the human mind to understand how certain AI systems operate), (ii) complexity, (iii) continuous adaptation and unpredictability, (iv) autonomous behaviour, and (v) data (functional dependence on data and the quality of data).

8 See Commission proposal for an AI act, Explanatory Memorandum and Recitals 1 and 5.

9 For the adoption of a harmonised set of requirements for AI systems.

10 For the adoption of specific rules for the processing of personal data in the context of biometric identification.

[11] The proposal complements both the sectoral product safety legislation, based on the new legislative framework (NLF) including the General Product Safety Directive, the Machinery Directive, the Medical Device Regulation and the EU framework on the approval and market surveillance of motor vehicles. The AI Act is also part of a broader EU regulatory framework comprising in addition the proposal for a new AI liability directive and the proposal for a revision of the product liability directive.

[12] See Article 2. The proposed regulation would also apply to the Union institutions, offices, bodies and agencies acting as a provider or user of AI systems.

[13] This covers the case of a service (digitally) provided by an AI system located outside the EU.

[14] See Council of Europe, Feasibility Study, Ad hoc Committee on Artificial Intelligence, CAHAI(2020)23 .

[15] OECD, Recommendation of the Council on Artificial Intelligence, 2019.

[16] See Article 3(1) and Recital 6.

[17] See impact assessment at pp. 48-49. A risk approach is also adopted in the United States Algorithmic Accountability Act of 2019 and in the 2019 Canadian Directive on Automated Decision-Making.

[18] FRTs would be allowed (i) for targeted search for potential victims of crime, including missing children, (ii) to prevent a specific, substantial and imminent threat to the life or physical safety of persons or of a terrorist attack, and (iii) for the detection, localisation, identification or prosecution of a perpetrator or individual suspected of a criminal offence referred to in the European Arrest Warrant Framework Decision.

[19] Harmonised standards are defined in accordance with Regulation (EU) No 1025/2012 and the Commission could, by means of implementing acts, adopt common technical specifications in areas where no harmonised standards exist or where there is a need to address specific safety or fundamental rights concerns.

[20] For an overview, see T. Madiega and H. Mildebrath, Regulating facial recognition in the EU, EPRS, September 2021.

[21] This section aims to provide a flavour of the debate and is not intended to be an exhaustive account of all different views on the proposal. Additional information can be found in related publications listed under 'EP supporting analysis'.

[22] For an in-depth analysis of the proposals and recommendations for amendments see N. Smuha and others, *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*, Elsevier, August 2021; M. Ebers, and others, *The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)*, J 4, no 4: 589-603, October 2021.

[23] N. Smuha, and others, above at pp. 14-15. See also E. Biber, Machines Learning the Rule of Law – EU Proposes the World's first Artificial Intelligence Act, August 2021. There are also calls for a shift in approach, to identify problematic practices that raise questions in terms of fundamental rights, rather than focusing on definitions; M. Veale and F. Zuiderveen Borgesius., Demystifying the draft EU AI Act, 22(4) *Computer Law Review International*, July 2021.

[24] See M. Ebers and others, above.

[25] See V. Galaz and others, Artificial intelligence, systemic risks, and sustainability, Vol 67, *Technology in Society*, 2021.

[26] For an overview, see T. Madiega and H. Mildebrath, above.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

© European Union, 2023.

eprs@ep.europa.eu (contact)

www.eprs.ep.parl.union.eu (intranet)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)

Second edition. The 'EU Legislation in Progress' briefings are updated at key stages throughout the legislative procedure.



Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 729.512 – June 2022

EN

Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

In recent years, the use of artificial intelligence (AI) in medicine and healthcare has been praised for the great promise it offers, but has also been at the centre of heated controversy. This study offers an overview of how AI can benefit future healthcare, in particular increasing the efficiency of clinicians, improving medical diagnosis and treatment, and optimising the allocation of human and technical resources.

The report identifies and clarifies the main clinical, social and ethical risks posed by AI in healthcare, more specifically: potential errors and patient harm; risk of bias and increased health inequalities; lack of transparency and trust; and vulnerability to hacking and data privacy breaches.

The study proposes mitigation measures and policy options to minimise these risks and maximise the benefits of medical AI, including multi-stakeholder engagement through the AI production lifetime, increased transparency and traceability, in-depth clinical validation of AI tools, and AI training and education for both clinicians and citizens.

AUTHORS

This study has been written by the following authors at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Karim Lekadir, University of Barcelona Department of Mathematics and Computer Science, Artificial Intelligence in Medicine Lab, Barcelona, Spain; Gianluca Quaglio, Panel for the Future of Science and Technology (STOA), European Parliament, Brussels, Belgium; Anna Tselioudis Garmendia, School of Public Health, Faculty of Medicine, Imperial College London, UK; Catherine Gallin, University of Barcelona Department of Mathematics and Computer Science, Artificial Intelligence in Medicine Lab, Barcelona, Spain.

ADMINISTRATOR RESPONSIBLE

Gianluca Quaglio, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in May 2022.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

Brussels © European Union, 2022.

PE 729.512
ISBN: 978-92-846-9456-3
doi:10.2861/568473
QA-07-22-328-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

Objectives

In recent years, a burgeoning interest in and concern over the use of artificial intelligence (AI) in medicine and healthcare has stood at the centre of interdisciplinary scientific research, political debate, and social activism. The goal of this report is to explain the areas in which AI can contribute to the medical and healthcare field, pinpoint the most significant risks relating to its application in this high-stakes and quickly-changing field, and present policy options to counteract these risks, in order to optimise the use of biomedical AI. Not only will this ensure the safety and respectful treatment of patients receiving AI-mediated healthcare, it should also aid the clinicians and developers involved in implementing it.

Methodology

This study employs an interdisciplinary methodology based on a comprehensive (but non-systematic) literature review and analysis of existing scientific articles, white papers, recent guidelines and regulations, governance proposals, AI studies, and online publications. The multi-disciplinary resources examined for this report include works from the fields of computer science, biomedical research, the social sciences, biomedical ethics, law, industry, and government reporting. This report explores a wide range of technical obstacles and solutions, clinical studies and results, as well as government proposals and consensus guidelines.

Specific applications of AI in medicine and healthcare

This study first outlines the potential for AI in medicine to address pressing issues, in particular the ageing population and the rise of chronic diseases, a lack of health personnel, inefficiency of health systems, lack of sustainability, and health inequities. The report also details the different fields in which biomedical AI could make the most significant contributions: 1) clinical practice, 2) biomedical research, 3) public health, and 4) health administration.

In the realm of clinical practice, the report goes into further detail concerning specific contributions – both realised and potential – to particular medical areas such as radiology, cardiology, digital pathology, emergency medicine, surgery, medical risk and disease prediction, adaptive interventions home care, and mental health. In biomedical research, the report details the potential contributions of AI to clinical research, drug discovery, clinical trials, and personalised medicine. Lastly, the report presents potential contributions of AI at the public health level as well as to global health.

Risks of AI in healthcare

This study identified and clarifies seven main risks of AI in medicine and healthcare: 1) patient harm due to AI errors, 2) the misuse of medical AI tools, 3) bias in AI and the perpetuation of existing inequities, 4) lack of transparency, 5) privacy and security issues, 6) gaps in accountability, and 7) obstacles in implementation. Each section, as summarised below, not only describes the risk at hand, but also proposes potential mitigation measures.

Patient harm due to AI errors

The study explains the main causes of AI errors: noise and artefacts in AI clinical inputs and measurements, data shift between AI training data and real-world data, and unexpected variations in clinical contexts and environments. The medical consequences of such errors may include missed

diagnosis of life-threatening conditions as well as false diagnosis, leading to inadequate treatment and incorrect scheduling or prioritisation of intervention.

Misuse of biomedical AI tools

AI tools, even when accurate and robust, are dependent on how human beings use them in practice and how the results they produce are used; in the healthcare context, these human actors include clinicians, healthcare professionals and patients. Incorrect usage of AI tools can result in incorrect medical assessment and decision making, and subsequently in potential harm for the patient.

Potential causes of AI misuse include limited involvement of clinicians and citizens in AI development, a lack of AI training in medical AI among healthcare professionals, lack of awareness and literacy among patients and the general public, and the proliferation of easily accessible online and mobile AI solutions without sufficient explanation and information.

Risk of bias in medical AI and perpetuation of inequities

Systemic human biases often make their way into AI models, including widespread and rooted bias based on sex and gender, race and ethnicity, age, socioeconomic status, geographic location, and urban or rural contexts. The most common causes of AI biases in the healthcare sphere are due to biased and imbalanced datasets which may be based on structural bias and discrimination (systemic discrimination that is imbedded in the ways that data is collected or the ways in which doctors treat their patients) and disparities in access to quality equipment and digital technologies, as well as lack of diversity and interdisciplinarity in technological, scientific, clinical, and policymaking teams.

Lack of transparency

A significant risk for AI is a lack of transparency concerning the design, development, evaluation, and deployment of AI tools. AI transparency is closely linked to the concepts of traceability and explainability, which correspond to two distinct levels at which transparency is required: 1) transparency of the AI development and usage processes (traceability), and 2) transparency of the AI decisions themselves (explainability).

Specific risks associated with a lack of transparency in biomedical AI include a lack of understanding and trust in predictions and decisions generated by the AI system, difficulties in independently reproducing and evaluating AI algorithms, difficulties in identifying the sources of AI errors and defining who and/or what is responsible for them, and a limited uptake of AI tools in clinical practice and in real-world settings.

Privacy and security

The increasingly widespread development of AI solutions and technology in healthcare, recently underscored by a reliance on big data during the Covid-19 pandemic, has highlighted the potential risks of a lack of data privacy, confidentiality and protection for patients and citizens. The main risks for data privacy and security in AI for healthcare, including personal data sharing without fully informed consent, data repurposing without the patient's knowledge, data breaches that could expose sensitive or personal information, and the risk of harmful – or even potentially fatal – cyberattacks on AI solutions, at both individual and hospital or health-system level.

Gaps in accountability

'Algorithmic accountability' is a crucial aspect of trustworthy and applicable AI in the field of healthcare. However, legal lacunae continue to exist in current national and international regulations concerning who should be held accountable or liable for errors or failures of AI systems, especially in medical AI. It is difficult to define the roles and responsibilities due to the multiplicity of actors involved in the process of medical AI, from design to deployment (e.g. healthcare professionals or AI developers). This lack of definition can leave clinicians and other healthcare

professionals in a particularly vulnerable position, especially if the AI model they are using is not entirely transparent.

Obstacles to implementation in real-world healthcare

Many medical AI tools have been developed recently; however, obstacles abound in the path towards implementation, integration and use of these tools in real-world clinical settings. Such obstacles include limited data quality, structure, and interoperability across heterogeneous clinical centres and electronic health records; potential alterations in the physician-patient relationship owing to the introduction of AI medical tools; increased and under-regulated access to patient data; and a lack of clinical and technical integration and interoperability of AI tools with existing clinical workflows and electronic health systems.

Risk assessment methodology

There is a need for a structured approach to risk assessment and management that specifically addresses the technical, clinical and ethical challenges of AI in healthcare and medicine.

Regulatory frameworks for AI

AI risks can be characterised and classified according to the severity of the harm they may induce, as well as to the probability and frequency of the harm induced. Currently, the applicable regulations for medical AI tools in the EU are the 2017/745 Medical Devices Regulation (MDR) and the 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR), which were passed in 2017. However, because they were derived at a time when AI was at an early stage in its development, many aspects specific to AI are not considered, such as continuous learning of AI models or the identification of algorithmic biases.

In 2021, the European Commission published a long-awaited proposal for an AI regulation and to harmonise the rules governing AI technologies across Europe. The highest category corresponds to AI tools that contradict EU values and hence should be prohibited. The intermediate category, which corresponds to high-risk AI and comprises medical AI technologies, can be permitted only when the tools comply with specific requirements and obligations for adequate risk management, such as ensuring human oversight and conducting post-market monitoring.

The European Commission proposal for AI regulation is general for all domains of society and does not take into account the specificities and risks of AI in the healthcare domain, contrary to the MDR and IVDR regulations. Furthermore, the European Commission proposal retains of some of the limitations of the MDR and IVDR, such as the lack of mechanisms to address the dynamic nature and continuous learning of medical AI technologies.

Risk minimisation through risk self-assessment

For risk identification in AI, several stakeholders have suggested a self-assessment, structured approach composed of specified checklists and questions. For example, the independent High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, published an assessment checklist for trustworthy AI called ALTAI. The checklist is structured around seven categories: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability.

The ALTAI model is general and does not address AI in healthcare specifically. This has motivated the recent development of consensus guidelines for trustworthy AI in medicine by a network of European Commission funded research projects together with international inter-disciplinary experts. Entitled FUTURE-AI, these guidelines are organised according to six principles (fairness, universality, traceability, usability, robustness, explainability) and comprise concrete

recommendations and a self-assessment checklist to enable AI designers, developers, evaluators and regulators to develop trustworthy and ethical AI solutions in medicine and healthcare.

Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions

While identifying and mitigating risks in medical AI by means of adequate evaluation studies is crucial, existing scientific literature focused mostly on evaluating model accuracy and robustness of the AI tools in laboratory settings. Other aspects of medical AI, such as clinical safety and effectiveness, fairness and non-discrimination, transparency and traceability, as well as privacy and security, are more challenging to evaluate in controlled environments and have thus received far less attention in scientific literature.

There is a need for a more holistic, multi-faceted evaluation approach for future AI solutions in healthcare. Best practices to enhance clinical evaluation and deployment include: (i) employing standard definitions of clinical tasks (e.g. disease definition) to enable objective community-driven evaluations; (ii) defining performance elements beyond accuracy, such as for fairness, usability, explainability and transparency; (iii) subdividing the evaluation process into stages of increasing complexity (i.e. to assess feasibility, then capability, effectiveness and durability); (iv) promoting external evaluations by independent third-party evaluators; and (v) employing standardised guidelines for reporting the AI evaluation results to increase reproducibility, transparency and trust.

Policy options

1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements

In order to tailor existing frameworks and AI practices specifically to the medical field, multi-faceted risk assessment should be an integral part of the medical AI development and certification process. Furthermore, risk assessment must be domain-specific, as the clinical and ethical risks differ in different medical fields (e.g. radiology or paediatrics). In the future regulatory framework, the validation of medical AI technologies should be harmonised and strengthened to assess and identify multi-faceted risks and limitations by evaluating not only model accuracy and robustness but also algorithmic fairness, clinical safety, clinical acceptance, transparency and traceability.

2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms

For the future acceptability and implementation of medical AI tools in the real world, many stakeholders beyond AI developers – such as clinicians, patients, social scientists, healthcare managers and AI regulators – will play an integral role. Hence, new approaches are needed to promote inclusive, multi-stakeholder engagement in medical AI and ensure the AI tools are designed, validated and implemented in full alignment with the diversity of real-world needs and contexts. Future AI algorithms should therefore be developed by AI manufacturers based on co-creation, i.e. through strong and continuous collaboration between AI developers and clinical end-users, as well as with other relevant experts such as biomedical ethicists.

Integrating human- and user-centred approaches throughout the whole AI development process will enable the design of AI algorithms that better reflect the needs and cultures of healthcare workers, while also enabling potential risks to be identified and addressed at an early stage.

3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI

New approaches and mechanisms are needed to enhance the transparency of AI algorithms throughout their lifecycle. From this need can emerge the concept of an 'AI passport' for standardised description and traceability of medical AI tools. Such a passport should describe and

monitor key information about the AI technology, covering at least five categories of information: 1) model-related information; 2) data-related information; 3) evaluation-related information; 4) usage-related information; and 5) maintenance-related information.

The AI passport should be standardised to enable consistent traceability across countries and healthcare organisations. Furthermore, the concept of traceability must go beyond the mere documentation of the development process or the phase of testing the AI model; instead, it should also comprise the process of monitoring and maintaining the AI model or system in the real world by continually tracking how it functions after deployment in clinical practice and identifying potential errors or changes in performance. Hence, it is important that algorithms are developed together with live interfaces that will be intended for continuous surveillance and auditing of the AI tools after their deployment in their respective clinical environments.

4. Develop frameworks to improve the definition of accountability and monitoring of responsibilities in medical AI

Frameworks and mechanisms are needed to assign responsibility adequately to all actors in the AI workflow in medical practice, including the manufacturers, thus providing incentives for applying all measures and best practices to minimise errors and harm to the patient. Such expectations are already an integral part of the development, evaluation and commercialisation of medicines, vaccines and medical equipment, and need to be extended to future medical AI products.

Another way to bolster accountability is through periodic audits and risk assessments, which can be used to evaluate how much regulatory oversight a certain AI tool might need. To this end, the assessments must be conducted through the whole AI pipeline, from data collection, to development, to pre-clinical stages, to deployment, but also when the tools are in use.

5. Introduce education programmes and campaigns to enhance the skills of healthcare professionals and the literacy of the general public in medical AI

To increase adoption and minimise error, future medical professionals should be adequately trained in medical AI, including its advantages in terms of improving care quality and access to healthcare, and its limitations and risks. It is therefore time to update educational programmes in medicine and increase their interdisciplinarity.

Furthermore, there is an urgent need to increase the AI literacy of the public so that citizens and patients can empower themselves and thus better take advantage of the benefits of emerging medical AI tools; increased AI literacy will also help minimise the potential risk of misuse of the AI tools, especially during remote monitoring and care management.

6. Promote further research on clinical, ethical and technical robustness in medical AI

There is a need for further research on the interrelated areas of medical AI to address the current clinical, socio-ethical and technical limitations. Examples of areas for future research include explainability and interpretability, bias estimation and mitigation, and secure and privacy-preserving AI.

More research is also needed to develop adaptation methods that can ensure a high level of generalisability of future AI tools across population groups, clinical centres and geographical locations. Future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions.

7. Implement a strategy for reducing the European divide in medical AI

While the EU has made significant investments in AI in recent years, inequalities persist between different European countries. The AI divide can be explained by structural differences in research programmes and technological capacities, as well as by varying levels of investment from the public and private sectors. The disparities in AI development and implementation between EU countries are particularly marked in medical AI. In this context, the EU can act as an umbrella to coordinate an EU-wide strategy for reducing the gaps in medical AI between European countries. This strategy should include concrete actions to boost the technological, research and industrial capacities of emerging EU countries in the field of AI for healthcare.

The EU Member States, in particular those in eastern Europe, could develop specific programmes to further support future AI in healthcare. The European Commission could implement specific coordination and support programmes of activities implemented in this sector by different Member States, thereby supporting the implementation of common guidelines and approaches. Furthermore, infrastructure projects should be established specifically for those EU countries that have limited research infrastructures and data availability. Existing education-focused programmes such as the Marie-Curie training networks could be strengthened to enhance training capacities and human capital in medical AI.

Table of contents

Executive summary I

1. Introduction 1

1.1. Objectives of this study 1

1.2. Methodology and resources used 1

1.3. Definitions 2

2. Artificial intelligence applications in healthcare 4

2.1. Artificial intelligence and healthcare needs 4

2.1.1. Main challenges for EU's healthcare systems 4

2.1.2. Main application domains for AI in healthcare 5

2.2. AI in clinical practice 5

2.2.1. Radiology 6

2.2.2. Digital pathology 6

2.2.3. Emergency medicine 6

2.2.4. Surgery 7

2.2.5. Risk prediction 7

2.2.6. Adaptive interventions 7

2.2.7. Home care 8

2.2.8. Cardiology 8

2.2.9. Nephrology 9

2.2.10. Hepatology 9

2.2.11. Mental health 10

2.3. AI in biomedical research 10

2.3.1. Clinical research 10

2.3.2. Drug discovery 11

2.3.3. Clinical trials 11

2.3.4. Personalised medicine 12
2.4. AI for public and global health 12
2.4.1. Public health 12
2.4.2. Global health 13
2.5. AI in healthcare administration 13
2.5.1. Coding 13
2.5.2. Scheduling 14
2.5.3. Detection of fraudulent activity 14
2.5.4. Patient flow management 14
2.5.5. Healthcare audits 14
3. Risk of AI in healthcare 15
3.1. Patient harm due to AI errors 15
3.2. Misuse of medical AI tools 17
3.3. Risk of bias in medical AI and perpetuation of inequities 20
3.4. Lack of transparency 22
3.5. Privacy and security issues 23
3.6. Gaps in AI accountability 25
3.7. Obstacles to implementation in real-world healthcare 27
4. Risk assessment methodology 30
4.1. Regulatory frameworks for AI 30
4.2. Risk minimisation through risk self-assessment 33
4.3. Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions 36
4.3.1. Standardised definition of clinical tasks 37
4.3.2. Multi-faceted evaluation of performance beyond accuracy 37
4.3.3. Subdivision of the evaluation process into discrete phases. 40
4.3.1. Promotion of external evaluations by third-party evaluators 42

4.3.2. Standardised and comprehensive reporting of the AI evaluation procedure and results ___ 43

5. Policy options ___ **46**

5.1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements ___ 46

5.2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms ___ 47

5.3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI ___ 48

5.4. Develop frameworks to better define accountability and monitor responsibilities in medical AI 49

5.5. Introduce education programmes to enhance the skills of healthcare professionals and the literacy of the general public ___ 50

5.6. Promote further research on clinical, ethical and technical robustness in medical AI ___ 51

5.7. Implement a strategy for reducing the European divide in medical AI ___ 52

References ___ **53**

List of figures

Figure 1 – Relationship between artificial intelligence, machine learning and deep learning ___ 3

Figure 2 – Main classes of AI tools reviewed in this report ___ 5

Figure 3 – Summary of causes and consequences of errors and failures of medical AI algorithms, together with some recommendations for potential mitigation ___ 16

Figure 4 – Main factors that can lead to incorrect use of medical AI algorithms by clinicians and citizens and potential mitigation measures to improve usability of future algorithms ___ 18

Figure 5 – Most common biases and their causes in medical AI, and potential mitigation measures to develop AI algorithms with increased fairness and equity ___ 20

Figure 6 – Main risks resulting from the current lack of transparency associated with AI algorithms followed by possible mitigation measures ___ 22

Figure 7 – Main privacy and security risks associated with big data and AI, and some mitigation measures ___ 24

Figure 8 – Current limitations in accountability and recommendations to fill in these gaps ___ 26

Figure 9 – Obstacles for clinical implementation and integration of new AI tools in real-world healthcare practice, together with potential mitigation measures ___ 28

Figure 10 – AI risk classification according to the 2021 EU proposal on AI legislation ___ 32

Figure 11 – Recommendations for improved evaluation of algorithm performance and risks in medical AI ___ 36

Figure 12 – Example of a multi-stage approach for medical AI evaluation ___ 41

Figure 13 – Summary of policy options suggested in this report ___ 46

Figure 14 – Example of a possible AI passport that can be used to improve traceability and transparency in medical AI, by documenting all key details about the AI tools, their intended use, model and data details, evaluation results, and information from continuous monitoring and auditing ___ 49

List of tables

Table 1 – Main definitions and concepts in medical AI _____ 2

Table 2 – Examples of performance elements for imaging AI algorithms _____ 38

Table 3 – Excerpts of subdivided evaluation process for medical AI, based on processes implemented in the drug development sector _____ 41

Table 4 – Reporting elements from the MINMAR reporting guidelines _____ 44

1. Introduction

1.1. Objectives of this study

In recent years, there has been growing interest in the application of artificial intelligence (AI) in healthcare. From drug discovery to healthcare provision, artificial intelligence (AI) has the potential to revolutionise the field of health. Precisely, AI will likely improve access to healthcare and how patients are treated, but it also optimises the way resources are allocated, thus helping health systems function more effectively and efficiently (EIT Health, 2020).

The potential for AI to reshape the field of healthcare – to help improve diagnosis and enable an increasingly personalised precision approach to medicine – may seem boundless. Some of the main applications of AI in medicine include medical image quantification, automated analysis of genetic data, disease prediction, medical robotics, telemedicine and virtual doctors. The coronavirus pandemic has accelerated the development and deployment of AI applications in the medical and clinical areas, as AI-related technologies lay at the main core of the response to this worldwide health crisis.

However, as with other technological advances, AI in the domain of healthcare comes with its specific benefits and risks, and needs its own set of regulatory frameworks that address the socio-ethical implications of its use. While the implementation of AI in healthcare holds great promise, this rapidly developing field also raises concerns for patients, healthcare systems and society; these concerns include issues of clinical safety, equitable access, privacy and security, appropriate use and users, as well as liability and regulation. Hence, researchers, the general public, and policymakers have all pointed to important bioethical issues, including how to evaluate the risks and benefits of AI in healthcare, how to establish accountability in the sphere of biomedical AI and how to regulate its use in this particularly high-stakes context. Another important question at the heart of the field is whether AI might increase inclusion and fairness in the treatment of traditionally underrepresented communities, or whether it runs the risk of perpetuating and augmenting pre-existing health disparities and inequities.

The study will provide an overview of AI health-related applications and an analysis of the potential of AI to transform the provision of healthcare. The study will also define, assess and clarify risks in the current and potential applications of AI in the domain of healthcare. At the same time, it will consider major clinical, socio-ethical and regulatory aspects of AI in its various health applications. Finally, the study will also propose a series of policy options aimed at minimising the risks of medical AI, enhancing governance at the EU level and strengthening its responsible development.

1.2. Methodology and resources used

The methodology implemented in this study is based on a comprehensive interdisciplinary (but non-systematic) literature review and analysis of existing scientific articles, white papers, recent guidelines, governance proposals, AI studies and results, news articles and online publications. These have been generated by AI developers, public agencies, expert leaders, clinical researchers, healthcare professionals and social scientists that have been actively working in the field of AI for medicine and healthcare in recent years, especially in the last two to three years.

A highly interdisciplinary body of literature was examined for this report, including works from the fields of computer science, biomedical research, the social sciences, biomedical ethics, law, industry, and government reporting. Hence, this report examines a wide range of technical obstacles and solutions, clinical studies and results, as well as government proposals and consensus guidelines.

A wide range of key phrase searches were performed in literature databases, in particular in Google Scholar, PubMed and Web of Science. Depending on the different themes investigated in this study, examples of key phrase searches include 'medical AI', 'AI risks', 'ethical challenges of AI', 'clinical safety', 'AI fairness', 'AI bias', 'AI inequities', 'AI accountability', 'data privacy', AI explainability', 'AI transparency', 'risk management', 'AI evaluation'.

In addition to summarising the considerations, findings and recommendations that apply to each of the themes examined in this report, concrete examples from a wide range of medical domains and applications (e.g. in radiology, cardiology, digital pathology, surgery, emergency medicine, etc.) are provided whenever possible to illustrate the challenges and potential future directions in medical AI.

1.3. Definitions

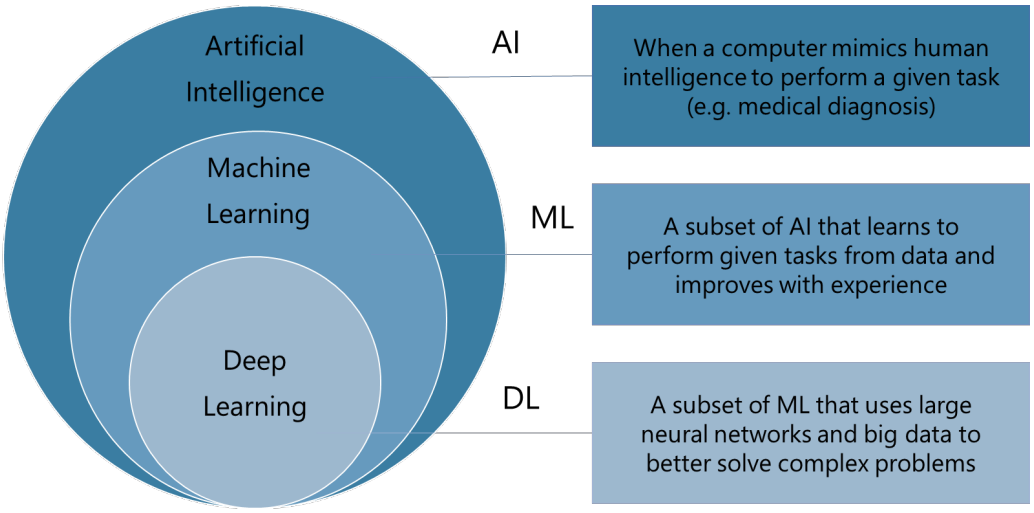
To introduce the readers to the field of AI, the table below provides a list of definitions of the main terms and concepts used throughout this report.

Table 1 – Main definitions and concepts in medical AI

Term	Definition
Artificial intelligence (AI)	Here we will first use the historical definition of AI, i.e. when a machine is able to mimic human intelligence or even surpass it to perform a given task such as prediction or reasoning. However, in this report, we will mostly focus on one subfield of AI that is dominant in the healthcare area, namely machine learning (ML).
Machine learning (ML)	ML is a subfield of AI and concerns the methods that learn to perform given tasks, such as prediction or classification, based on existing data.
Big data	The term big data is used in instances in which the data samples are too large to be adequately analysed with traditional AI methods. In this case, new methods such as deep neural networks (otherwise known as deep learning) can be used (Raghupathi et al., 2014).
Neural networks (NNs)	NNs, technically known as artificial NNs, are circuits composed of a set number of interconnected neurons organised hierarchically in layers and which are capable of learning to perform highly complex tasks from data. Each neuron acts as a type of specialised processing unit which transforms input data into output signals. These transformations are application specific and learned from available application-specific data. Progressively, the neurons combine their outputs, layer by layer, approximating the processing of a large complex function, until the network outputs a final result, such as the prediction of a disease (Esteva et al., 2019).
Deep learning	DL refers to NNs with more than three layers; in this case, the availability of big data is needed to estimate the optimal values of the parameters for this larger, more complex type of deep neural network (Goodfellow et al., 2016). Note that not all AI and ML tools are based on deep learning or NNs. Other techniques such as decision trees or support vector machines are widely used, especially when the data sample is not sufficiently large to build NNs or deep NNs (Figure 1).
AI model, AI algorithm or AI tool	Technically, in the specialised AI literature, an AI algorithm is the procedure used to build an AI model for a specific application, hence the AI model is the output of the machine learning algorithm. In other words, the same AI algorithm can be used to build models (e.g. predictive models) for many different applications, but

	the AI model is specific to a given application (e.g. predicting the patient's response to a given cancer treatment). However, the terms AI algorithms and AI models (or ML algorithms and ML models) are often used interchangeably. AI tools are AI models that are packaged to be used by end-users, so they contain more than just the AI model, such as user interfaces. In non-specialised literature, AI models, algorithms, tools, solutions and software are used interchangeably, especially in medical circles.
Training, validation & testing data	Training data are datasets that are used by AI developers to train their AI models. Validation data are also used by AI developers. However, the latter is used to optimise the parameters of the AI models so that they can be applied to new data other than the training data. In other words, validation data are used to fine-tune the AI models to make them generalisable (to use a terminology from the technical literature). Testing data are new data that are distinct from those used for training and optimising the AI models. They are used to evaluate the AI models, ideally by evaluators that did not take part in the AI development phase (in other words by external independent evaluators, though in practice AI models are still widely evaluated by the same teams that developed them in the first place).
Medical AI or healthcare AI	This is a type of AI which is focused on specific applications in medicine or healthcare.
AI design, development, evaluation & deployment	These are roughly the main steps of the AI lifecycle in healthcare. First the AI tools are designed, generally in a co-creation approach and through collaborations between AI developers and clinical experts in the field (and sometimes by also involving patients and other experts such as healthcare managers). The AI developers write some code to build and optimise the AI models from the training and validation data they have at their disposal. Subsequently, the AI model is evaluated using testing data that is distinct from the training and validation data. The AI tool (AI model with a user interface) is also evaluated with end-users (e.g. doctors and/or patients). If the evaluation is successful and convincing for the relevant stakeholders (e.g. patients, clinicians, healthcare managers, regulatory authorities), the AI tool is validated, approved, and then deployed in practice. The forementioned pipeline is of course an ideal scenario, and in practice there is some degree of variation in the AI development lifecycle.

Figure 1 – Relationship between artificial intelligence, machine learning and deep learning



2. Artificial intelligence applications in healthcare

Information generated by medical science currently spans a very wide scope; it is rapidly growing and will continue to do so both in volume and variety. In parallel, the potential for AI in medicine and health is massive and is constantly expanding as AI technologies are being developed by industry, academia, government, and individuals. It is expected that the integration of AI-based technologies into medical practice will produce substantial changes in many areas of medicine and healthcare (Roski et al., 2019; Fihn et al., 2019).

2.1. Artificial intelligence and healthcare needs

2.1.1. Main challenges for EU's healthcare systems

Before reviewing the most recent developments in medical AI in this chapter, it is important to first detail the main healthcare challenges and unmet needs that could benefit from the deployment of AI in future medical care:

Ageing population and chronic diseases. In 2017, approximately 37% of the ageing population of the EU member states reported having at least two chronic diseases, on average. Among people aged 80 and over, 56% of women and 47% of men reported multiple chronic diseases on average across EU countries (OECD/European Union, 2020).

Lack of health personnel. European countries suffer from gaps in the supply and skill level of health personnel. An estimated overall shortfall of 1.6 million healthcare workers in the EU was reported in 2013; in order to compensate for this shortage, an annual exponential growth greater than 2% would be needed. However, as this rate of increase has not been reached, the expected shortage is anticipated to reach 4.1 million by 2030 (0.6 million physicians, 2.3 million nurses and 1.3 million other healthcare professionals) (WHO, 2016; Michel, 2020).

Inefficiency. There is ample evidence of widespread inefficiency in EU healthcare systems (OECD, 2017). While the relative ability of a particular healthcare system to transform resources into outcomes differs across countries, there is considerable waste of health-related resources, which contributes to excessive expenditure (Medeiros, 2015).

Sustainability. The issue relating to health-systems sustainability is rapidly growing in the EU. According to the OECD 'Health at a glance: Europe 2020' report, the EU spends 8.3% of its GDP on healthcare, with marked differences in spending across regions: in Germany and France, it is 11% and in Luxembourg and Romania, less than 6%. Health expenditure is projected to continue to escalate, mainly due to sociodemographic changes – the ageing population and the subsequent increase in chronic diseases and long-term care needs – as well as the impact of new technologies. In addition to the aforementioned challenges, in recent years EU healthcare systems have also been under significant pressure due to economic difficulties (Quaglio, 2020). The COVID-19 pandemic in particular is expected to increase the health spending share of GDP in multiple countries.

Healthcare inequities. Healthcare inequities and inequalities persist among the EU member states and their populations. The right of every EU citizen to timely access to affordable, preventive, and curative care of high quality is one of the key principles of the newly proclaimed European Pillar of Social Rights (European Commission. The European Pillar, 2021). A recent report identified several challenges and inequalities related to healthcare access, namely: (a) inadequate public resources invested in the healthcare system; (b) fragmented population coverage; (c) gaps in the range of benefits covered; (d) prohibitive user charges, in particular for pharmaceutical products; (e) lack of protection of vulnerable groups from user charges; (f) lack of transparency on how waiting list priorities are set; (g) inadequate availability of services, particularly in rural areas; (h) problems with

attracting and retaining health professionals; (i) difficulties in reaching particularly vulnerable communities who have limited access to qualitative healthcare such as ethnic minorities and socioeconomically disadvantaged people; (j) racial bias and unequal healthcare provision (European Commission. A study of national policies 2018; Hamed, 2020).

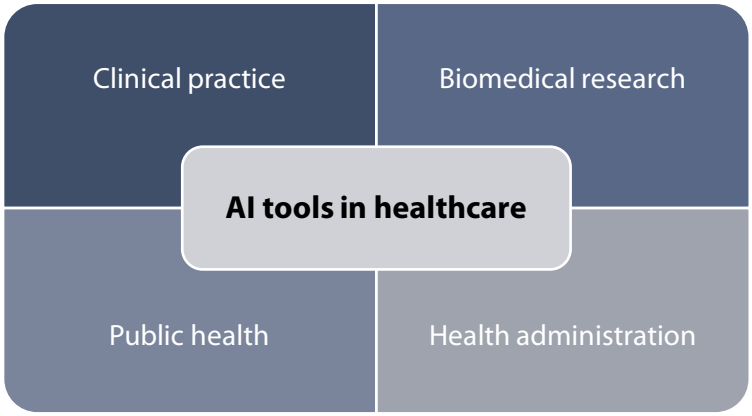
2.1.2. Main application domains for AI in healthcare

To date, AI has progressively been developed and introduced into virtually all areas of medicine, from primary care to rare diseases, emergency medicine, biomedical research and public health. Many management aspects related to health administration (e.g. increased efficiency, quality control, fraud reduction) and policy are also expected to benefit from new AI-mediated tools (Gómez-González, 2020).

Healthcare AI tools have been often classified according to the stakeholder user groups, i.e. 1) patients and citizens; 2) clinicians and caregivers; 3) healthcare administrators; and 4) public health professionals and policy makers. Classification of biomedical AI tools can also be based on the setting in which the tools are used: 1) clinical settings (hospitals, primary care centres, emergency care centres); 2) clinical processing and managing settings (laboratory, pharmacy, radiology, etc); and 3) administrative settings.

For the purpose of this study, we adopt a more comprehensive classification of AI applications, dividing them into four practices: 1) clinical; 2) research; 3) public health; and 4) administrative (Figure 2). The next sections provide a summary of the current developments and applications of AI in these four areas.

Figure 2 – Main classes of AI tools reviewed in this report



2.2. AI in clinical practice

The potential for the application of AI in the clinical setting is enormous and ranges from the automation of diagnostic processes to therapeutic decision making and clinical research. The data necessary for diagnosis and treatment comes from many sources, including clinical notes, laboratory tests, pharmacy data, medical imaging, and genomic information.

AI will play a major role in tasks such as automating image analysis (e.g. radiology, ophthalmology, dermatology, and pathology) and signal processing (e.g. electrocardiogram, audiology, and electroencephalography). In addition to its implementation in test and image interpretation, AI could be used to integrate and array results with other clinical data to facilitate clinical workflows (Topol et al., 2019). Many impressive examples exist in clinical settings where AI tools are applied, a number of which are expounded below. The following sections also touch on the possible application of AI into specific areas of medicine that are more scarcely reported, such as nephrology and personalised medicine.

2.2.1. Radiology

Radiology is among the medical specialities that have seen significant AI developments over the last years. Imaging AI technologies show promise in assisting radiologists in the work of medical image quantification. For example, segmentation with limited human supervision has been achieved by using deep network models, which enable to automatically localise and delineate the boundaries of anatomical structures or lesions (Peng & Wang, 2021). These AI tools can also prioritise and track findings that mandate early attention, and enable radiologists to concentrate on images that are most likely to be abnormal (Lee et al., 2018; Peng & Wang, 2021). A good example of AI tools for medical image segmentation is 'cvi42', a cardiovascular imaging platform commercialised by the Canadian company Circle CVI that has been adopted in over 40 countries (Zange et al., 2019).

Radiomics is another imaging processing technique in which AI has proven useful. Although the term is not strictly defined, radiomics generally aims to extract quantitative information (the so-called radiomic features), from diagnostic and treatment planning images (Gillies, 2016; Mayerhoefer et al., 2020). Radiomic features capture tissue and lesion characteristics, such as heterogeneity and shape, and may be used for clinical problem solving alone or in combination with demographic, histologic, genomic, or proteomic data. The impact of radiomics increases when the wealth of information that it provides is processed using AI techniques (Cook et al., 2019; Mayerhoefer et al., 2020).

A recent meta-analysis compared the performances of deep learning software and radiologists in the field of imaging-based diagnosis (Liu, 2019). According to the study, the diagnostic performance of deep learning models is equivalent to that of healthcare professionals. However, a major finding of the review is that most of the studies analysed have serious limitations: (i) most studies took the approach of assessing deep learning diagnostic accuracy in isolation (many studies were excluded at screening because they did not provide comparisons between the human and the machine); (ii) very few studies reported comparisons with health professionals using the same test dataset; (iii) there were very few prospective studies done in real clinical environments (most studies were retrospective and based on previously assembled datasets); iv) the scrutinised studies showed inconsistencies over key terminology.

2.2.2. Digital pathology

The term digital pathology was initially coined to include the process of digitising whole-slide images using advanced slide-scanning techniques. It now also refers to AI-based approaches for the detection and analysis of digitised images (Bera et al., 2019; Niazi et al., 2019). While the use of standardised guidelines can support the harmonisation of diagnostic processes, histopathological analysis is inherently limited by its subjective nature and by differences in judgement between independent experts (Chi et al., 2016; Evans et al., 2008; Bera et al., 2019).

AI can contribute to the alleviation of some of the challenges faced by oncologists and pathologists, including inter-subject and inter-operator variability. Several studies demonstrate that AI can have a similar level of accuracy to that of pathologists (Ehteshami Bejnordi et al., 2017) and, more significantly, can improve their diagnostic performances when used in tandem (Steiner et al., 2018; Bera et al., 2019). In digital pathology, AI has been applied to a variety of image processing and classification tasks. These include low-level tasks such as detection, focused on object recognition problems (Sornapudi et al., 2018), as well as higher-level tasks such as predicting disease diagnosis and prognosis (Corredor et al., 2018), evaluating disease severity and outcome (Mobadersany et al., 2018) and using assays to predict response to therapy (Bera, 2019).

2.2.3. Emergency medicine

Emergency medicine can benefit from AI in different phases of patient management. For instance, it offers potential value for improved patient prioritisation during triage, and is versatile in analysing

different elements of the patient's clinical history. Currently, patients are assessed with limited information in the emergency department (Berlyand et al., 2019; Kirubarajan et al., 2020). However, there is potential for emergency department flow metrics and resource allocation to be optimised through AI-driven decision making (Berlyand et al., 2018). Nevertheless, concerns remain regarding the use of AI for patient safety considering the limited body of evidence to support its implementation (Challen et al., 2019; Kirubarajan et al., 2020).

A recent scoping review analysed the applications of AI in emergency medicine in a total of 150 studies (Kirubarajan et al., 2020). According to the review, the majority of interventions are centred on: (i) the predictive capabilities of AI; (ii) improving diagnosis within the emergency department; (iii) studies focused on triage of emergent conditions; and iv) studies demonstrating that AI can assist with organisational planning and management within the emergency department.

2.2.4. Surgery

In the area of surgery, decisions sometimes need to be taken under time constraints and conditions of uncertainty regarding an individual patient's diagnoses and predicted response to treatment. Uncertainty may be imposed by unavailability of patient data (e.g. external hospital records or diagnostic tests) or absence of high-level evidence to guide important management decisions. Under such time constraints and uncertainty, clinicians may instead rely on cognitive shortcuts and snap judgments using pattern recognition and intuition (Dijksterhuis et al., 2006; Loftus et al., 2020).

Ultimately, these factors can lead to bias, error and preventable harm. In a number of conditions, traditional decision-support tools appear not to be sufficiently equipped to accommodate time constraints and uncertainty regarding diagnoses and the predicted response to treatment, both of which can impair surgical decision making (Loftus, 2020). These challenges can be overcome by AI models (Loftus et al., 2019). In fact, AI tools provide diverse sources of information (patient risk factors, anatomic information, etc.) that can help in the development of better surgical decisions (Shickel et al., 2019; Hashimoto et al., 2019).

2.2.5. Risk prediction

Risk prediction focuses on assessing the likelihood of individuals experiencing a specific health condition or outcomes. It typically generates probabilities for a wide array of outcomes ranging from death to adverse disease events (e.g. stroke, myocardial infarction, bone fracture). The process involves the identification of individuals with certain diseases or conditions and their classification according to stage, severity, and other characteristics. These individuals may subsequently be targeted to receive specific medical interventions (Miotto et al., 2016; Steele et al., 2018; Fihn et al., 2019).

Risk prediction models have long been available in healthcare. However, these are currently based on regression analysis and subsets of available clinical data, resulting in limited prediction accuracy which renders them less valuable in the clinical setting. Importantly, the advent of large repositories of data and AI techniques has shown promising signs for AI's usefulness in tailoring patient-specific conventional approaches for risk prediction (Islam, 2019). For example, predictive AI-based models in cardiovascular disease risk assessments have shown improved performance when compared to statistically derived predictive risk models (Jamthikar et al., 2019).

2.2.6. Adaptive interventions

Adaptive interventions, also defined as 'just-in-time adaptive interventions', are intervention designs aimed to deliver the right type and level of support by continuously adapting to an individual's changing internal and contextual states (Almirall et al., 2014). In particular, this allows to adjust the frequency, duration and dosage of medicines at different time points throughout the course of care.

AI-driven adaptive interventions can provide support in medical treatment through two different pathways: (i) direct input, via self-assessments by patients; or (ii) via passive data collection, where physiological information is gathered using special sensors. Using mobile technologies to collect self-assessments is referred to as ecological momentary assessment (De Vries et al., 2020). The later helps people to self-monitor behaviours at the time and in the context in which they occur.

For example, ecological momentary assessment has several benefits in substance–use disorders, such as increasing the ability to correlate instances of craving with maladaptive behaviours. Passive data collection often relies on technologies that record patterns of movement within the patient's environment, for example, via global positioning system (GPS) and wireless local area networks (Wi-Fi), which are used to acquire location-based data (Vijayan et al., 2021).

The possibility to gather spatial and temporal information (i.e. where and when the behaviours of the subject occurred) renders these tools highly specific. In addition, physiological information from special sensors (such as those measuring blood pressure, heart rate, temperature or substance concentration levels in blood), can be combined with spatial and temporal data in order to get a more detailed profile of the patient's behaviour, including monitoring physiological responses or precursors to craving (Quaglio et al., 2019).

2.2.7. Home care

In 2019, more than one fifth (20.3%) of the EU-27 population was aged 65 and over. The share of people aged 80 years or above is projected to have a two and a half folds increase between 2019 and 2100, from 5.8% to 14.6% (Eurostat. Statistical expanded, 2020). It is worth noting that the prevalence of dementia increases rapidly with age (Quaglio et al., 2016). In 2018, an estimated 9.1 million people aged over 60 were living with dementia in EU Member States (around 7% of the population aged over 60), compared to 5.9 million in 2000. In fact, the percentage of people living with dementia in EU countries is expected to rise by about 60% over the next two decades and reach 14.3 million by 2040 (OECD/EU, 2018).

Importantly, AI can play a significant role in the self-management of chronic diseases and diseases that affect the elderly. Self-management tasks range from taking medications to adjusting the patient's diet and managing health devices. Home monitoring has the potential to increase independence and improve ageing at home by keeping track of physical space and falls. In particular, tools, software, smartphone and mobile applications can enable patients to manage a large part of their own healthcare and facilitate their interactions with the healthcare system (Sapci et al., 2019).

Nevertheless, smart homes present several inconveniences, namely: 1) changing the lifestyle of users; 2) difficulties in the use of smart home technologies; 3) interoperability between systems; and 4) privacy and security constraints. Despite the current advances, the adoption of these emerging home-based technologies still falls short of end-user needs, prompting the search for new strategies (Azzi et al., 2020).

2.2.8. Cardiology

The most promising application of AI is for the automated processing of cardiac imaging data, which is necessary for the assessment of cardiac structure and function in cardiology (Lopez-Jimenez et al., 2020). Cardiac imaging modalities such as cardiac ultrasound, cardiac computer tomography and cardiovascular magnetic resonance imaging provide complex spatiotemporal data that are tedious and time consuming to process by cardiologists. The availability of new AI-driven cardiac image processing techniques has revolutionised cardiac clinical practice by enabling cardiologists to make more rapid assessment of the patients in their day-to-day practice (Lopez-Jimenez et al., 2020).

Machine learning (ML) models are set to improve the diagnostic capacity of echocardiography which constitutes the predominant cardiac imaging modality but remains heavily reliant on human expertise (Alsharqi et al., 2018). The generation of more accurate and automated echocardiograms with the use of AI is expected to reveal unrecognised imaging features that will facilitate the diagnosis of cardiovascular disease while minimising the limitations associated with human interpretation.

This is already the case in electrocardiography (ECG), for which AI models – such as deep-learning convolutional neural networks – have been generated with the use of large digital ECG datasets derived from clinical records (Siontis et al., 2021). As a result, AI-enabled ECGs are now capable of identifying diseases such as asymptomatic left ventricular dysfunction and silent atrial fibrillation, as well as phenotypic features including sex, age and race (Adedinsewo et al, 2020; Attia et al., 2019a; Attia et al., 2019b; Noseworthy et al., 2020).

Furthermore, AI has been used extensively in nuclear cardiology, which studies non-invasive imaging tools evaluating myocardial blood flow, among other things. ML models have been applied to two techniques in particular; single-photon emission computed tomography (SPECT) and myocardial perfusion imaging (MPI), to ultimately enhance the detection and prognosis of obstructive coronary artery disease (Noseworthy et al., 2020). It is believed that cardiac risk scores (calculating the 10-year risk of presenting with cardiovascular disease) will be assessed more accurately with the use of ML algorithms capable of extrapolating information and delineating unseen patterns in data derived from clinical records (Quer et al., 2021).

Although cardiovascular medicine appears to be at the forefront of AI in health, it will always, to a certain extent, depend on the expertise of cardiovascular specialists. Therefore, it is important for practitioners to be actively involved in this new and emerging field in order for imaging processing techniques to reach their full potential and perhaps revolutionise patient care (Quer et al., 2021).

2.2.9. Nephrology

The application of AI in nephrology is more scarcely reported than in other fields of medicine (Lindenmeyer et al., 2021; Chaudhuri et al., 2021). Nevertheless, its potential is increasingly being recognised by clinicians due to the promising advances made in the last decade. For instance, a novel deep learning model for ultrasound kidney imaging non-invasively classifies chronic kidney disease (CKD) (Kuo et al., 2019). In addition, the digital analysis of histopathological images has been facilitated by the development of a deep neural network capable of annotating and classifying human kidney biopsies (Hermsen, 2019). In an attempt to ameliorate early treatment of acute kidney injury (AKI), scientists took advantage of the widespread increase in data found in electronic healthcare records to develop an AI model enabling up to 48h prediction of inpatient episodes of AKI (Tomašev, 2019). On the other hand, the so-called 'Intraoperative Data Embedded Analytics' (*IDEA*) algorithm has been trained to predict the risk of developing postoperative AKI by integrating physiological data derived before and after an operation (Adhikari et al., 2019).

AI also holds potential in the computer-aided diagnosis of kidney cancer. As algorithms are becoming more robust and generalisable, they are increasingly better at identifying renal masses and distinguishing between benign and cancerous ones (Giulietti et al., 2021). Overall, the implementation of AI models in nephrology will likely facilitate prognosis, reinforce personalised medicine and reduce the global burden of kidney diseases (Park et al., 2021).

2.2.10. Hepatology

AI research is steadily progressing in many areas of medicine, and hepatology is no exception (Ahn et al., 2021). ML models have been used extensively to facilitate the diagnosis of multiple types of liver disease, most of which are life threatening. Interest has been primarily focussed on the automated detection of non-alcoholic fatty liver disease (NAFLD), as most patients remain

asymptomatic until the development of liver cirrhosis. A recently developed AI neural network shows 97.2% accuracy in diagnosing NAFLD (Okanoue et al., 2021).

Importantly, the same model is capable of distinguishing between patients with NAFLD and those with its more advanced form, NASH (non-alcoholic steato-hepatitis). Predictive models have also been developed to estimate the severity and prognosis of chronic viral hepatitis, as well as acute-on-chronic liver failure (Ahn et al., 2021). Despite the considerable progress in AI and hepatology, a number of conditions remain under-researched in this aspect, such as alcohol-associated liver disease and genetic/autoimmune liver disease, which calls for a more widespread adoption of AI in hepatology (Ahn et al., 2021).

2.2.11. Mental health

The EU suffers from a significant mental health burden. Neuropsychiatric disorders constitute 26% of diseases in EU Member States. Up to 40% of years lived with disability in the EU can be attributed to these types of mental health disorders, and especially to depression (WHO, 2021a). The cost of mood disorders and anxiety in the EU is about €170 billion per year (WHO, 2021a). In addition, it has been shown that depression and anxiety contribute greatly to chronic sick leave from the workplace and that these disorders – especially major depression – are often left untreated.

There is potential for AI to lend support to mental health patients and to mitigate the effects of a paucity of health personnel dedicated to mental health conditions. In fact, various tools are currently under development. These include digital tracking of depression and mood via keyboard interaction, speech, voice, facial recognition, sensors, and the use of interactive chatbots (Firth et al, 2017; Fitzpatrick et al., 2017; Mohr et al., 2018).

The computational power harnessed by AI systems could be leveraged to reveal the complex pathophysiology of psychiatric disorders and thus better inform therapeutic applications (Graham 2019; Lee, 2021). Machine learning has been explored to predict the efficacy of antidepressant medication (Chekroud et al., 2016), characterising depression (Wager et al., 2017), predicting suicide (Walsh et al., 2017) and psychosis in schizophrenics (Chung et al., 2018).

AI can help to differentiate between diagnoses with overlapping clinical presentations but with different treatment options (Dwyer et al., 2018). Examples include the identification of bipolar versus unipolar depression (Redlich et al., 2014), or the differentiation between types of dementia (Lee et al., 2021).

Nowadays, social media represent a form of daily communication for an extensive part of the population. Therefore, examining the content and language patterns of social media can provide insights and create new opportunities for predictive psychiatric diagnosis. Mental conditions may become observable in online contexts, while social media information analysed with machine learning has already been leveraged to predict diagnoses and relapses (Reece et al., 2017; Birnbaum et al., 2019; Yazdavar et al., 2020; Lee et al., 2021).

2.3. AI in biomedical research

2.3.1. Clinical research

Biomedical research seems to benefit more from AI-derived solutions compared to clinical applications, with recent advances also showing promising applications of AI in clinical knowledge retrieval. For example, mainstream medical knowledge resources are already using ML algorithms to rank search results, including algorithms that continuously learn from users' search behaviour (Fiorini et al., 2018a).

One example is PubMed, a widely used search engine for biomedical literature (Fiorini et al., 2018b). The AI technologies implemented by PubMed to optimise its search function include machine learning and natural language processing algorithms that are trained on patterns found in users' activities in order to improve a user's search (Fiorini et al., 2018b). For instance, Best Match is a new search algorithm for PubMed that leverages the intelligence of PubMed users and cutting-edge ML technology as an alternative to the traditional date sort order. The Best Match algorithm is trained using past user searches with dozens of relevance-ranking signals (factors), with the most important being the past usage of an article, publication date, relevance score, and type of article. This algorithm has significantly improved the finding of relevant information over the default time order in PubMed and has increased usage of relevance search over time (Fiorini et al., 2018b). Through techniques such as information extraction, automatic summarisation, and deep learning, AI has the potential to transform static narrative articles into patient-specific clinical evidence (Elliott et al., 2014).

2.3.2. Drug discovery

Drug designers frequently apply ML techniques to extract chemical information from large compound databases and to design new drugs. Central to this shift is the development of AI approaches to implement innovative modelling based on the large nature of drug datasets. As a result, recently developed AI approaches provide new solutions to enhance the efficacy and safety evaluation of candidate drugs based on big data modelling and analysis.

AI models such as these can facilitate greater understanding of a wide range of types of drugs and the clinical outcomes that they may offer (Zhu et al., 2020). For example, researchers recently trained a deep learning algorithm to predict molecules' potential antimicrobial activity. The algorithm screened over one billion molecules and virtually tested over 107 million, identifying eight antibacterial compounds that were structurally distant from known antibiotics (Stokes et al., 2020).

Compared to traditional animal models, both in vitro and in silico testing have great potential in lowering the cost of drug discovery. The application of in vitro and in silico approaches in the early stages of drug research and development procedures can reduce the number of drug attritions (Zhang et al., 2017). AI holds great potential as a method to assess compounds according to their biological capacities and toxicities. Existing AI models, such as those based on quantitative structure-activity relationship (QSAR) approaches (Golbraikh et al., 2016), can be used to predict large numbers of new compounds for various biological end points.

However, the resulting QSAR model predictions of new compounds are characterised by a number of limitations (Zhao et al., 2017; Zhu et al., 2020). Over the past decade, new efforts have stimulated the development of high-throughput screening (HTS) techniques (Zhu et al., 2014). HTS is a process that screens thousands to millions of compounds using standardised protocol. Facilitated by the combined efforts of HTS and combinatorial chemical synthesis, modern screening programmes can produce enormous amounts of biological data (Zhu et al., 2020).

2.3.3. Clinical trials

Randomised controlled trials (RCT) are the most robust method of assessing the risks and benefits of any medical intervention. However, undertaking an RCT is not always feasible. Common difficulties of unsuccessful RCTs include poor patient selection, inadequate randomisation, insufficient sample size, and poor selection of end points (Lee et al., 2020). AI models can be trained to better select the study participants with advanced statistical methods, and to assess study end points in a data-driven method. The application of AI will generate more efficient execution and greater statistical power than the one expected from traditional RCTs (Lee et al., 2020).

In addition to the efficient selection process, having a sufficiently large sample size is critical to enable detection of statistically significant differences between groups. Many RCTs require a considerable sample size because the effect of the treatment in question can be small. AI has the potential to select the right patients for RCTs. Furthermore, AI may enable more sensitive quantification of key study end points compared to the way they are usually measured. AI will also improve and complement RCTs significantly in the future. However, enhanced collaboration and synergy among clinicians, researchers, and industries is required for AI algorithms to be used to their full potential in RCTs (Lee et al., 2020).

2.3.4. Personalised medicine

Personalised medicine strongly relies on a scientific understanding of how an individual patient's unique characteristics, such as molecular and genetic profiles, make this patient vulnerable to a disease and sensitive to a therapeutic treatment (Strianese et al., 2020). Hundreds of genes have been identified for their contributions to human illness, and genetic variability in patients has also been used to distinguish individual responses to treatments (Zhu et al., 2020; Strianese et al., 2020).

The original concept of personalised medicine has been expanded to include other properties and individual clinical characteristics to ultimately form a new concept called 'extended personalised medicine'. The latter is developed from additional sources of information such as clinical sources, demographic data, social data, lifestyle parameters (sleep hours, physical activity, nutritional habits, etc), environmental conditions, etc. (Gómez-González, 2020).

AI tools may enhance the progress made in personalised medicine by evaluating the clinical benefit of different research methods and multiple data types (Mamoshina et al., 2018). Drug-target predictions (Sydow et al., 2019), metabolic network modelling, and population genetics pattern identifications (Schrider et al., 2018) constitute some of the recent advancements in this field that rely on computational modelling (Lorkowski et al., 2021). To truly impact routine care, however, the data needs to represent the diversity of patient populations (OECD, 2020). Therefore, the shift toward a data-driven personalised medicine system will have far-reaching implications for patients, clinicians, and the pharmaceutical industry (Boniolo et al., 2021).

2.4. AI for public and global health

2.4.1. Public health

Public health has many definitions, but one that is frequently used is that it is 'the science and art of preventing disease, prolonging life and promoting health through the organised efforts and informed choices of society, organisations, public and private, communities and individuals' (Wanless, 2004). Experiments with relevant AI solutions are currently under way within a number of public health areas. A selected number of these areas are discussed below.

AI can help identify specific demographics or geographical locations where the prevalence of disease or high-risk behaviours exist (Maharana & Nsoesie, 2018; Shin et al., 2018). The range of AI solutions that can improve disease surveillance is also considerable. Digital epidemiological surveillance refers to the integration of case- and event-based surveillance (e.g., news and online media, sensors, digital traces, mobile devices, social media, microbiological labs, and clinical reporting) to analyse approaches for threat verification. This has been implemented to build early warning systems for adverse drug events and air pollution (Mooney & Pejaver, 2018).

AI has already made inroads into environmental and occupational health through data generated by sensors and robots. AI has the potential to intensify contact with patients, as well as to target services to patients. An essential component of these initiatives involves contacting large numbers

of patients via a variety of automated, readily scalable methods, such as text messaging and patient portals (Fihn et al., 2019).

2.4.2. Global health

AI may provide opportunities to address health challenges in low-and middle-income countries (LMICs). These challenges include acute health workforce shortages and weak public health surveillance systems. Although not unique to such countries, these challenges are particularly relevant in low- and middle-income settings, given their contribution to morbidity and mortality (Schwalbe & Wahl, 2020). For example, in some instances, AI-driven interventions have supplemented clinical decision making towards reducing the workload of health workers (Guo & Li, 2018). New developments in AI have also helped identify disease outbreaks earlier than traditional approaches (Lake et al., 2019).

AI studies in LMICs have also addressed public health from a broader perspective: more specifically in health policy and management. These studies include AI research aimed at improving the performance of health facilities, improving resource allocation from a systems perspective, and reducing traffic-related injuries in addition to other health system issues (Schwalbe & Wahl, 2020).

Although AI can help in addressing several existing and emerging health challenges in LMICs, many issues warrant further exploration. These issues relate to the development of specific AI-driven health interventions and their real efficacy and effectiveness. Additionally, ethical regulatory standards should be implemented in order to help protect the interests and needs of the local communities and attempt to increase community-based research and engagement (Collins et al., 2019). Finally, the successful deployment of many AI tools in LMICs will require investment to strengthen the underlying healthcare systems (Schwalbe & Wahl, 2020).

2.5. AI in healthcare administration

Healthcare systems are characterised by a heavy administrative workflow with a wide range of actors and institutions, comprising patients (e.g. management of billing), health professionals, healthcare facilities and organisations (e.g. patient flow), imaging facilities, laboratories (e.g. supply chain of consumables), pharmacies, payers, and regulators. A report carried out in a primary care setting identified several potential areas of concern within this heavy administrative setting. These include time spent on reclaiming financial reimbursement, entering data into various unintegrated practice-based information systems, processing information from hospitals and other external providers and helping patients navigate a fragmented health system. The study concluded that over 50% of practice time was spent on bureaucracy, the majority of which was potentially avoidable (Clay & Stern, 2015).

AI can perform these routine tasks in a more efficient, accurate and unbiased fashion. One argument in favour of using AI in administrative practices is that errors in these activities are less serious than errors in the clinical setting. However, the danger of hacking, lack of privacy and security remains (Roski et al., 2019; OECD, 2020). AI applications can be critical in the organisation of patient flow. For example, lack of bed availability is an important cause of surgical cancellations (Kaddoum et al., 2016); however, it is a preventable administrative error in patient flow. This problem occurs frequently and is also associated with delays in discharge in clinical ward (Stylianou et al., 2017).

2.5.1. Coding

Coding is the process of extracting information from clinical records and codifying it using classifications such as the International Classification of Diseases (ICD) or diagnosis-related groups (DRGs). Coding is a complex, labour-intensive process, and coding accuracy is very important for reimbursement, administration and research. While computer-assisted coding has existed for more

than a decade, AI can enhance the accuracy and transparency of this administrative practice (OECD, 2020).

2.5.2. Scheduling

Scheduling is another example in which AI can add value to the administrative process. Algorithms fed on historical data can predict which patients may not attend their appointments, allowing practitioners to take proactive action to manage the situation. Beyond blanket or even targeted reminders, AI can address a patient's needs and queries (OECD, 2020).

2.5.3. Detection of fraudulent activity

Algorithms can also learn to look for fraudulent activity in healthcare, i.e. using a code for a more expensive medical service than the one performed (OECD, 2020).

2.5.4. Patient flow management

The fluent management and transfer of patients through the different stages of care with minimal delays is what defines patient flow (NHS, 2017). Notably, the quality of the services provided by the healthcare systems as well as patient satisfaction should be maintained throughout. Poor patient flow has been shown to negatively affect patients, staff, and the overall quality of care (Tlapa et al., 2020). Technological solutions such as AI are increasingly applied to purposes associated with patient flow (Dawoodbhoy et al., 2021). For example, the fluctuating volume of patient arrivals is a crucial but uncertain variable in hospital emergency departments.

Knowing the patient arrival volume in advance enables the smooth operational planning of emergency departments and improves related decision making (Menke et al., 2014; Ram et al., 2015). By implementing better resource planning and allocation based on predictive outcomes, the probability of overcrowding can be reduced to ultimately improve healthcare quality (Jiang et al., 2018).

2.5.5. Healthcare audits

Healthcare auditing is the process of reviewing patients' records in order to identify recommendations for improvement (NHS England, 2021). This process provides both quantitative information on the current state of affairs as well as recommendations on how to improve clinical outcomes. Audits can be carried out routinely or in the instance of a significant shortcoming in the delivery of a service, such as an increase in infection rates (Nagar et al., 2015) or patient flow concerns (Kamat & Parker, 2015).

3. Risk of AI in healthcare

In an article published more than 50 years ago, William B. Schwartz stated that 'computing science will probably exert its major effects by augmenting and, in some cases, largely replacing the intellectual functions of the physician' (Schwartz, 1970). Despite promising examples of healthcare AI solutions, Schwartz's prediction has not yet been fully realised. Initial results of AI health applications are not as robust as predicted and it is difficult to assess their real impact (Roski et al., 2019; Fihn et al., 2019).

Some players claim that the potential of AI medicine as a whole has been largely overestimated, with virtually no data demonstrating an actual improvement in patient outcomes (Angus, 2020; Parikh, 2019; Emanuel, 2019). Other experts have raised concerns over the last years regarding potential adverse consequences of medical AI, including clinical, technical and socio-ethical risks (Challen et al., 2019; Gerke & Cohen, 2020; Ellahham et al., 2020; Morley & Floridi, 2020; Manne & Kantheti, 2021).

In this chapter, we will describe the main risks that have been identified in the literature as likely to arise from the introduction of AI in future healthcare. We will focus on seven categories of risks and challenges:

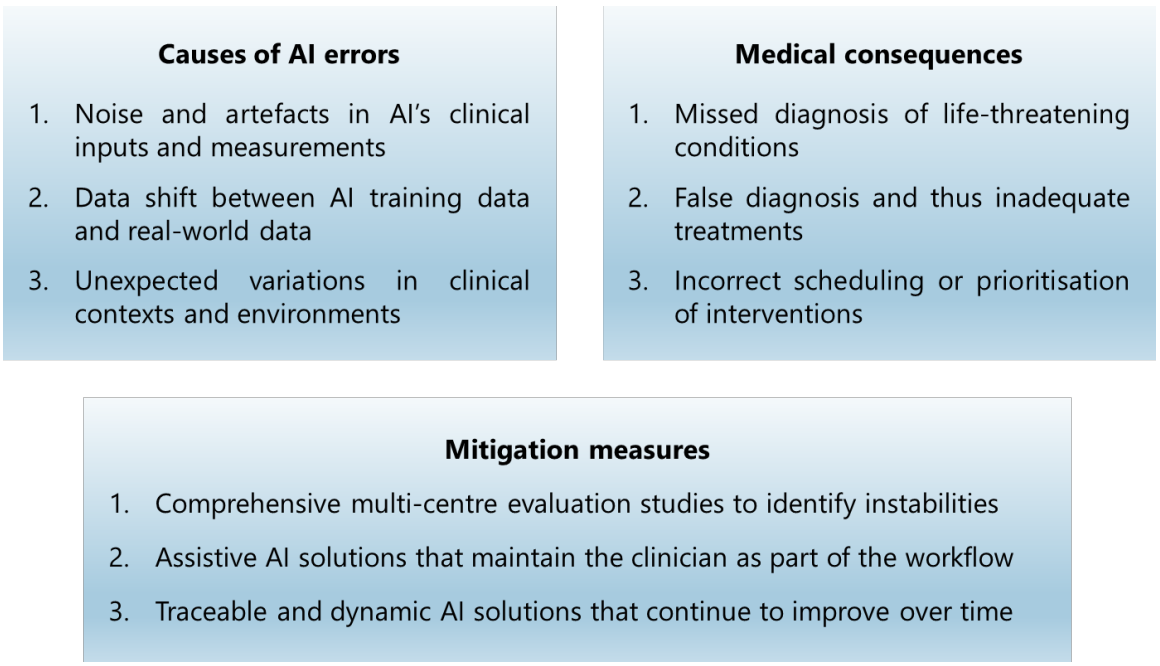
1. Patient harm due to AI errors
2. Misuse of medical AI tools
3. Risk of bias in medical AI and perpetuation of inequities
4. Lack of transparency
5. Privacy and security issues
6. Gaps in AI accountability
7. Obstacles to implementation in real-world healthcare

Not only could these risks result in harms for the patients and citizens, but they could also reduce the level of trust in AI algorithms on the part of clinicians and society at large. Hence, risk assessment, classification and management must be an integral part of the AI development, evaluation and deployment processes.

3.1. Patient harm due to AI errors

Despite continuous advances in data availability and machine learning, AI-guided clinical solutions in healthcare may be associated with failures that could potentially result in safety concerns for the end-users of healthcare services (Challen et al., 2019; Ellahham et al., 2020). These AI algorithm errors can lead, for example, to (1) false negatives in the form of missed diagnoses of life-threatening diseases, (2) unnecessary treatments due to false positives (healthy persons incorrectly classified as diseased by the AI algorithm), (3) unsuitable interventions due to imprecise diagnosis, or incorrect prioritisation of interventions in emergency departments (Figure 3).

Figure 3 – Summary of causes and consequences of errors and failures of medical AI algorithms, together with some recommendations for potential mitigation



Assuming that AI developers have access to large-scale datasets with sufficient quality for training their AI technologies, there are still at least three major sources of error for AI in clinical practice. Firstly, AI predictions can be significantly impacted by noise in the input data during the usage of the AI tool. For example, ultrasound scanning – the most commonly used imaging modality in clinical practice due to its low-cost and portability – is known to be prone to scanning errors (Farina et al, 2012). This depends particularly on the experience of the operator, the cooperation of the patient, and the clinical context (e.g. emergency ultrasound) (Pinto et al., 2013). Even in high-income countries where there is a high level of medical training, such errors are expected to occur in some scans, thus affecting subsequent AI predictions.

Secondly, AI misclassifications may appear due to dataset shift (Subbaswamy et al., 2020), a common problem in machine learning that occurs when the statistical distribution of the data used in clinical practice is shifted, even slightly, from the original distribution of the dataset used to train the AI algorithm. This shift could be due to differences in the population groups, acquisition protocols between hospitals, or the usage of machines from different manufacturers. A recent study (Campello et al., 2020) has shown that AI models trained on cardiac magnetic resonance image (MRI) scans from two scanners (e.g. Siemens and Philips) lose accuracy when applied to MRI data acquired from different machines (e.g. General Electric and Canon).

Another example of dataset shift can be seen in a multi-centre study in the United States that built a highly accurate pneumonia diagnosis AI system based on data from two hospitals (Zech et al., 2018). When tested with data from a third hospital, a significant decrease in accuracy was noticed, suggesting potential hospital-specific biases. In another example, the company DeepMind developed a deep learning model trained on a large dataset for automated diagnosis of retinal diseases from optical coherence tomography (OCT) (De Fauw, et al., 2018). They found that the AI system was confused when applied to images obtained from a machine that is different from the one used for data acquisition at the AI training stage, with the diagnosis error increasing from 5.5% to a staggering 46%. These examples illustrate the current challenges posed in building AI tools that maintain a high level of accuracy even if the data is heterogeneous across populations, hospitals or machines.

Lastly, the predictions can be erroneous due to the difficulty of AI algorithms to adapt to unexpected changes in the environment and context in which they are applied. To illustrate the problem, researchers at Harvard Medical School described a nice example in the domain of AI for medical imaging (Yu & Kohane, 2019). They imagined an AI system that was trained to detect shadows or dense features on a chest X-ray images that are associated with lesions in major diseases such as lung cancer. Then, they listed a number of simple scenarios in which the AI may lead to incorrect predictions, such as if the X-ray technician leaves the adhesive ECG connectors on their patient's chest or if the patient wears a wedding ring and places their hand on their chest during the scan. In these scenarios, it is possible that the AI model could mistake these circular artefacts as one of the known chest lesions, resulting in a false positive.

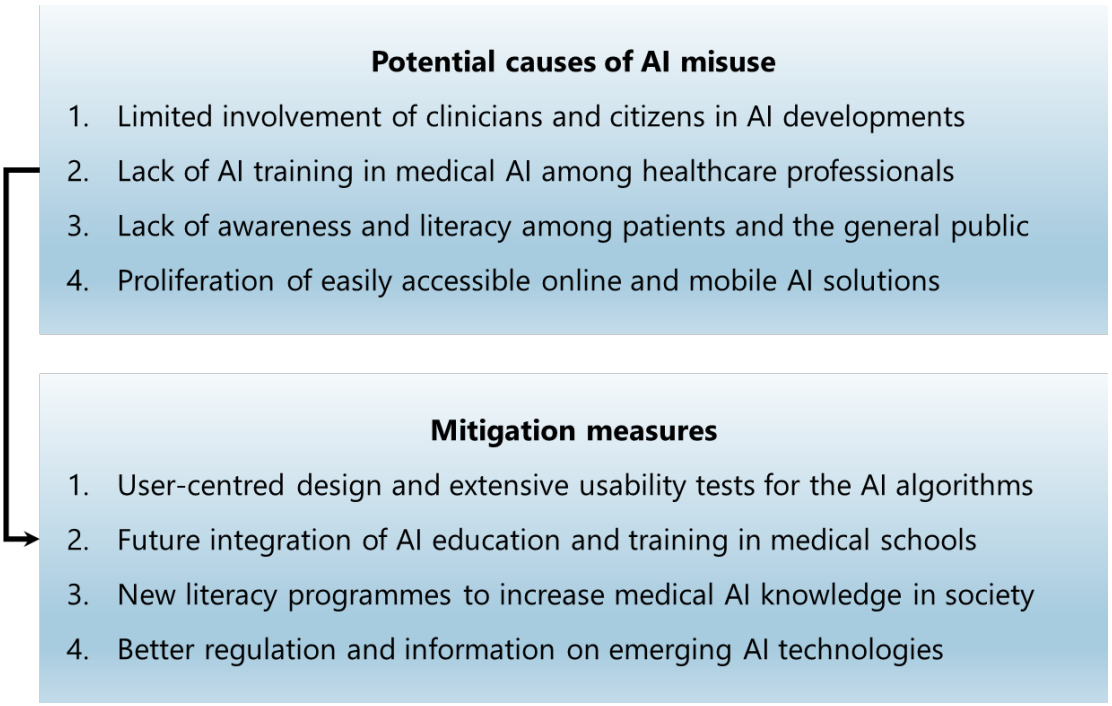
There are at least three avenues to minimise the risk of AI errors and safety issues for patients (Figure 3). First of all, standardised methods and procedures need to be defined for extensive evaluation and regulatory approval of AI solutions, in particular regarding their generalisability to new populations and sensitivity to noise. Second, the AI algorithms should be designed and implemented as assistive tools (as opposed to fully autonomous tools), such that clinicians remain part of the data processing workflow to detect and report potential errors and contextual changes, and hence to minimise harm to patients.

Furthermore, future AI solutions in healthcare must be dynamic, i.e., they should be embedded with mechanisms to continue to learn from new scenarios and mistakes as they are detected in practice. However, this last aspect will still require a certain degree of human control and vigilance to identify problems as they appear; this in turn may increase costs and reduce the initial benefits of AI. Infrastructural and technical developments will also be needed to enable regular AI updates (based on past and new training), and it will be necessary to implement policies that ensure such mechanisms are integrated into healthcare settings.

3.2. Misuse of medical AI tools

As with most health technologies, there is a risk for human error and human misuse with medical AI. Even when the developed AI algorithms are accurate and robust, they are dependent on the way they are used in practice by the end-users, including clinicians, healthcare professionals, and patients. Incorrect usage of AI tools can result in incorrect medical assessment and decision making and subsequently in potential harm for the patient. Hence, it is not enough for clinicians and the general public to have access to medical AI tools, but it is also necessary for them to understand how and when to use these technologies.

Figure 4 – Main factors that can lead to incorrect use of medical AI algorithms by clinicians and citizens and potential mitigation measures to improve usability of future algorithms



There are multiple factors that make existing medical AI technologies prone to human error or incorrect use (Figure 4). First, they have often been designed and developed by computer/data scientists with limited involvement from end-users and clinical experts. As a result, it is the user (i.e., the clinician, the nurse, the data manager or the patient) that is required to learn to use and to adapt to the new AI technology, which can lead to unnatural and complex interactions and experiences. In turn, the clinical user may encounter difficulties in understanding and applying the AI algorithm in day-to-day practice, which will limit the perception of informed decision making, while increasing the chances of human error.

This problem is exacerbated by the fact that existing training programmes in medicine are not yet tailored for medical AI and generally do not equip new clinicians with knowledge and skills in the area of AI. A survey performed in Australia and New Zealand in 2021 with 632 medical trainees (in the areas of ophthalmology, dermatology, and oncology) showed that 71% of the respondents believed AI would improve their field of medicine, especially for improved disease screening and streamlining of monotonous tasks (Scheetz et al., 2021).

However, most respondents indicated that they had never used AI applications in their work as a clinician (>80%) and only 5% viewed themselves as having excellent knowledge of the field. Another study performed in the United Kingdom surveyed 484 students from 19 medical schools and found that none of the students received any AI teaching as part of their compulsory curriculum (Sit et al., 2020). Similar conclusions were reached on knowledge and utilisation of technology-based interventions among health professionals in the European Union in other healthcare domains (Quaglio et al., 2019).

These reflections on AI education and literacy also apply to citizens and patients, who will become active users of future medical AI solutions. A 2021 study performed in five countries (Australia, the United States, Canada, Germany, and the United Kingdom) with over 6,000 citizens showed that the public generally has low awareness and understanding of AI and its use in everyday life (Gillespie et al., 2021). While younger people, men, and the university-educated tend to be more aware and

understand AI better, even these groups report low to moderate AI understanding (Gillespie et al., 2021).

Another cause for potential misuse of medical AI, which could lead to harm for citizens and patients, is the proliferation of easily accessible medical AI applications. For example, commercial mobile apps have been developed by several companies for skin cancer detection with the purpose of enabling individuals to take and upload a picture of their skin through the app, which is then directly analysed and assessed by the app's AI algorithm. Some examples of such apps include Skinvision, MelApp, skinScan and SpotMole.

While these tools are easily accessible to the general public, there is often limited information on how the AI algorithms in question have been developed and validated, while their reliability and clinical efficacy is not always demonstrated. For example, a recent study which evaluated six mobile apps for skin cancer detection demonstrated their lack of efficiency and high risk for bias (Freeman et al., 2020). The authors concluded: '*Current algorithm-based smartphone apps cannot be relied on to detect all cases of melanoma or other skin cancers. The current regulatory process for awarding the CE marking for algorithm-based apps does not provide adequate protection to the public*' (Freeman et al., 2020).

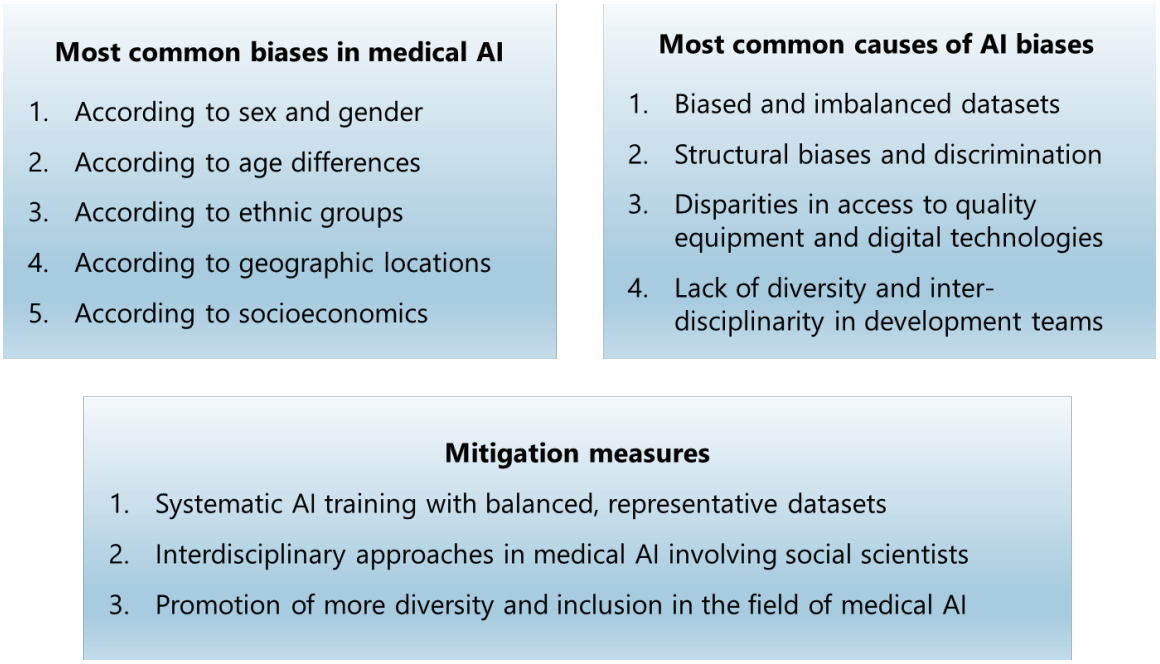
A quick search shows that many AI-powered online/mobile tools have also emerged in a wide range of medical domains and are commercially offered for medical diagnostics and health monitoring, such as Diagnostics.ai, DDXRX Doctor Ai, Symptomate, and Achu Health. While such services can constitute a promising solution for remote diagnosis and disease follow-up, their wide proliferation online can become a public health concern, in the same way that easily accessible online pharmacies have contributed to an abuse of medication by citizens (Bandivadekar , 2020).

Since there is a lot of financial gain to be made from the development and commercialisation of AI-powered web/mobile health applications, this sector will continue to attract a lot of new players and companies with varying standards of ethics, excellence and quality. The companies offering these web or mobile based AI medical tools acknowledge on their websites that their AI products are not certified medical devices and the terms of service often contain disclaimers. One can easily find disclaimers such as '*this site is designed to offer you general health information for educational purposes only*' or '*the health information furnished on this site and the interactive responses are not intended to be professional advice and are not intended to replace personal consultation with a qualified physician, pharmacist or other healthcare professional*'. However, most users may not necessarily come across, read and comprehend these disclaimers, and hence may rely on potentially incorrect information and diagnoses provided by the AI tools, which may negatively impact their decision making regarding their health.

There are several avenues to reduce human error or incorrect use of future medical AI solutions (Figure 4). First of all, end-users such as healthcare professionals, specialists, technicians or patients should be closely involved in the design and development of AI solutions to ensure their points of view, preferences and contexts are well integrated into the final tools that will be deployed and used. Furthermore, education and literacy programmes on AI and medical AI should be developed and generalised across education circles and society to increase the knowledge and skills of future AI end-users and hence reduce human error. Finally, it is important that public agencies help regulate the sector of web/mobile medical AI, such that the citizens are well informed and protected against the misuse and abuse of these emerging, easily accessible AI technologies.

3.3. Risk of bias in medical AI and perpetuation of inequities

Figure 5 – Most common biases and their causes in medical AI, and potential mitigation measures to develop AI algorithms with increased fairness and equity



Despite continuous advances in medical research and healthcare delivery, there remain important inequalities and inequities in medical care within most countries around the world. The main factors that contribute to these inequalities and inequities include sex/gender, age, ethnicity, income, education and geography. While some of these inequities are systemic, such as due to socioeconomic differences and discrimination, human biases also play an important role. For example, in the United States, existing research has demonstrated that doctors do not take Black patients' complaints of pain as seriously nor do they respond to them as quickly as they do for their White counterparts (Hoffman et al., 2016). Persistent in most countries around the world, to varying degrees, is yet another example of common bias embedded in healthcare systems: gender-based discrimination. Once again, in the domain of pain management, studies have pointed to the increased psychologisation or invisibilisation of female patients when reporting pain (Samulowitz et al., 2018).

Hence, in the recent years, there have been concerns that, if not properly implemented, evaluated and regulated, future AI solutions could embed and even amplify the systemic disparities and human biases that contribute to healthcare inequities. A few examples of algorithmic biases have already made the headlines in recent years, some of which are detailed below.

A study published in *Science* in 2019 showed that an algorithm used in the United States to help in the referral process of patients who need extra or specialist care was shown to discriminate against Black patients (Obermeyer et al., 2019). The authors of the study explained that with the algorithm, *'at a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%'.* A Canadian study in 2020 evaluated the degree of fairness of state-of-the-art deep learning algorithms used to detect abnormalities such as fractures, lung lesions, nodules, pneumonia, etc. in chest X-ray images (Seyyed-Kalantari et al., 2020). The study showed that the highest rate of underdiagnosis was in young females (age: 0-20), in Black patients, and in patients on public health insurance for low-income people and households. Furthermore, patients with intersectional identities (for example, a Hispanic female patient on low-

income health insurance) suffered the highest rates of underdiagnosis. The authors concluded that '*models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modification*' (Seyyed-Kalantari et al., 2020).

It is widely argued that the most common cause for unfairness in medical AI is the bias in the data used to train the machine learning models. As Marzyeh Ghassemi from the University of Toronto stated in a recent presentation on AI in healthcare (Ghassemi, 2021): '*Bias is already part of the clinical landscape. So, it is not as if machine learning is out to get us. It is that when we are training on data that humans make, that humans label, that humans annotate, we might pick up on some of the biases that humans have injected into that data*'.

As an example, in 2002 the National Lung Screening Trial, which compiled datasets from 53,000 smokers to investigate methods for early diagnosis of lung cancer, was found to include only 4% of Black participants in the data (Ferryman & Pitcan, 2018). Machine learning algorithms for skin cancer detection have been all-too-often trained on highly biased datasets – such as the International Skin Imaging Collaboration, one of the most widely used open-access database of skin lesions – which contain images from mostly fair-skinned patients in the United States, Europe, and Australia (Adamson & Smith, 2018). Diagnostic models only trained on fair-skin groups could prove to be detrimental to the diagnostic process of melanoma lesions present on dark-skinned individuals. Similarly, the way COVID-19 appears to affect patients differently according to their sex group means an AI algorithm trained on existing clinical data is likely to suffer from reduced fairness when predicting severity and mortality in men and women (Jin et al., 2020).

Another type of bias that appears in datasets is of a geographic nature. In 2020, researchers from the fields of radiology and biomedical research at Stanford University conducted a review of articles published over a five-year period that had been used in training deep learning algorithms related to patient care (Kaushal et al., 2020). They found that 71% of the United States studies in which geographic location was identified used data only from California, Massachusetts, and New York. In addition, they found the studies did not include any data from 34 of the 50 states in the U.S. Geographic bias can be an important issue in Europe too, as data availability and access to digital equipment are unevenly distributed, particularly in the Eastern European regions (EGA Consortium, 2021).

Another potential source of lack of fairness in medical AI is bias in the data labelling during clinical assessment. For example, existing research has shown that due to gender stereotypes, women are over-diagnosed for some diseases such as depression and under-diagnosed for other diseases such as cancer (Dusenberry, 2018). Furthermore, a large-scale Danish study, which analysed data on hospital admissions for approximately 7 million citizens and 19 disease groups, found that for the vast majority of the diseases, women are diagnosed later than men (Westergaard et al., 2019). Importantly, for many of these medical conditions such as injury, poisoning, congenital malformations and infectious diseases, these discrepancies cannot be explained by anatomical or genetic differences. If the data labels in the health registries are affected by such healthcare disparities, such as in environments where given groups have been systematically misdiagnosed due to stigma or stereotypes, then the AI models will likely learn to perpetuate this disparity (Rajkomar et al., 2018).

In recent years, awareness of algorithmic bias has increased and researchers, particularly in North America, have started to investigate mitigation measures to address the risk of unfairness in medical AI. First, it is evident that AI developers, in collaboration with clinical experts and healthcare professionals, must pay close and continuous attention to the selection and labelling of the data and variables to be used during model training. These should be representative and balanced with respect to key attributes such as sex/gender, age, socioeconomics, ethnicity, as well as geographic location. Furthermore, it is recommended to involve not only data scientists and biomedical researchers in the development teams, but also social scientists, biomedical ethicists, public health

experts, as well as patients and citizens. The latter group must be as diverse as possible to ensure that adequate diversity of backgrounds, experiences and needs are taken into consideration during the AI production lifecycle and that the tools created are truly representative and founded on community-based research.

3.4. Lack of transparency

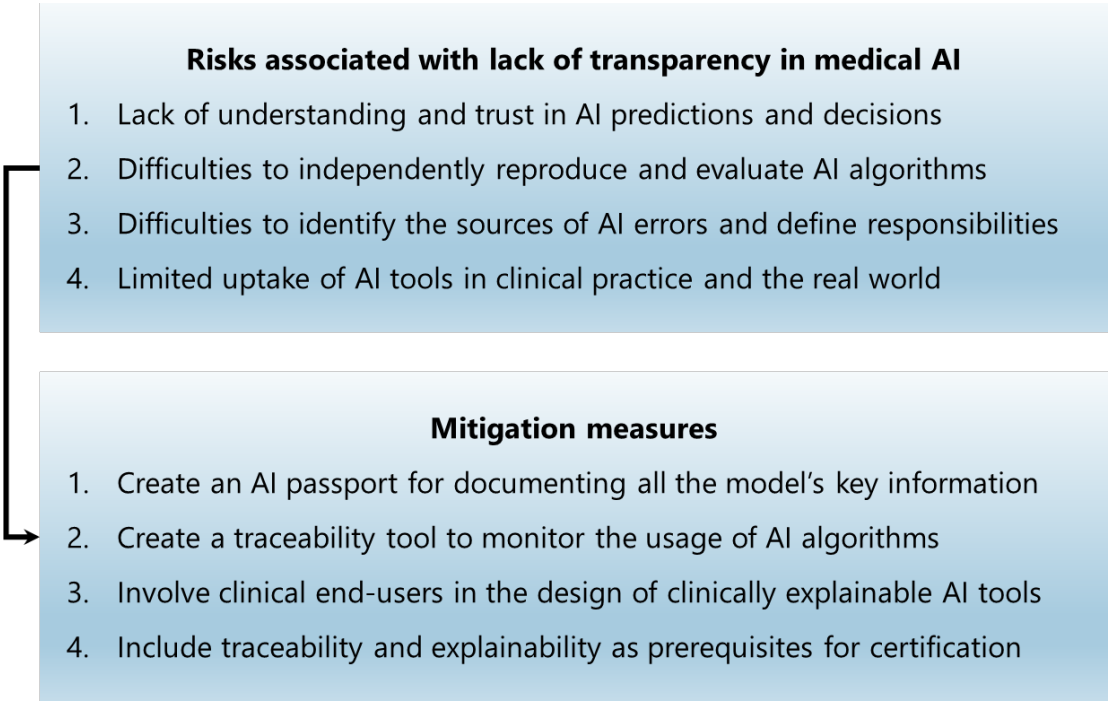
Despite continuous advances in medical AI, existing algorithms continue to be viewed by individuals and experts alike as complex and obscure technologies, which are difficult to fully comprehend, trust and adopt.

A recent AI algorithm developed by Google for breast cancer screening received considerable attention for its promising performance (McKinney , 2020): It was shown to improve the speed and robustness of breast cancer screening, to generalise well to populations in multiple countries beyond those used for training, and it even outperformed radiologists in specific situations. However, this work also received some criticism in the media and in the AI community as it was presented with almost no details on how the algorithm was built and on key technical descriptions. Some critics questioned the usefulness and safety of such an AI tool (Wiggers, 2020; iNews, 2020), while a group of scientists used this algorithm as their central example when they published a call in Nature for more transparency in medical AI (Haibe-Kains et al., 2020).

Lack of transparency is widely regarded as an important issue in the development and use of current AI tools in healthcare (Figure 6). It is expected to result in a great lack of trustworthiness in AI especially in sensitive areas such as medicine and healthcare that are focused on the wellbeing and health of citizens. At the same time, a lack of trustworthiness will evidently impact the level of adoption of emerging AI algorithms by patients, clinicians, and healthcare systems.

AI transparency is closely linked to the concepts of traceability and explainability, which correspond to two distinct levels at which transparency is required, i.e. (1) transparency of the AI development and usage processes (traceability), and (2) transparency of the AI decisions (explainability).

Figure 6 – Main risks resulting from the current lack of transparency associated with AI algorithms followed by possible mitigation measures



Traceability is considered a key requirement for trustworthy AI, and refers to transparently documenting the whole AI development process, including tracking how the AI model functions in real-world practice after deployment (Mora-Cantallops et al., 2021). More specifically, traceability requires maintaining a complete account of (i) model details (intended use, type of algorithm or neural network, hyper-parameters, as well as pre- and post-processing steps), (ii) training and validation data (gathering process, data composition, acquisition protocols and data labelling) and (iii) AI tool monitoring (performance metrics, failures, periodic evaluations) (EU Regulation, 2017; FDA, 2019).

In practice, existing AI tools in healthcare are rarely delivered with full traceability. In fact, companies often prefer not to disclose too much information about their algorithms, which are thus delivered as opaque tools that are difficult to understand and examine by independent parties. This, in turn, reduces the level of trust and adoption into real-world practice.

While traceability addresses the transparency of the AI algorithm's lifecycle, AI explainability is important for providing transparency for each AI prediction and decision. Article 22 of the European Union's General Data Protection Regulation (GDPR) details the 'right to explanation' which requires an explanation to be offered regarding the automated decision-making process (Selbst & Powles, 2017).

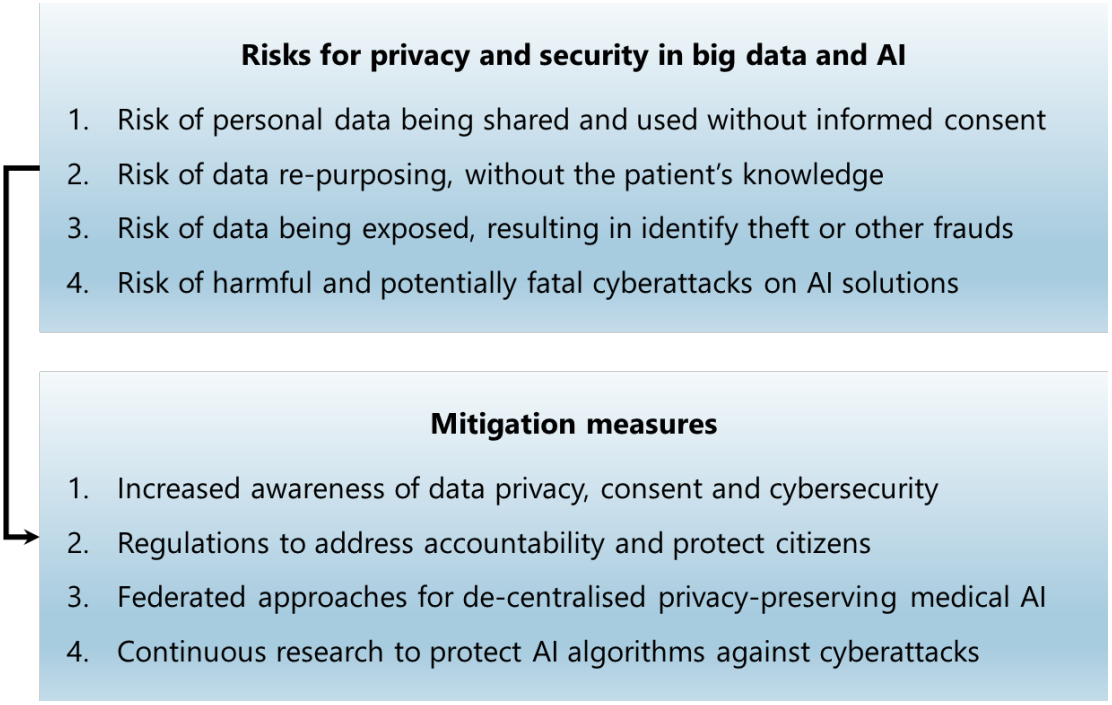
However, AI solutions, and specifically deep neural networks lack transparency, and are often described as 'black box AI', referring to the fact that these models learn complex functions that humans struggle to understand (Yang et al., 2021) and whose functions and decision-making processes are not visible or understandable. A lack of transparency makes it difficult for clinicians and other stakeholders to incorporate AI solutions into their real-world practice because in order to work with specific AI solutions, clinicians need to be able to understand the fundamental principles behind each decision and/or prediction, even when the algorithm itself has the potential to enhance the clinician's productivity (Lipton, 2017). Furthermore, the lack of explainability means that it is difficult to identify the source of AI errors and define responsibilities when it goes wrong.

There are numerous avenues available to improve the transparency of AI technologies in healthcare. First of all, there is a need for an 'AI passport' that could be a requirement for each AI algorithm for documenting all the model's key information. There is also a need to develop traceability tools for monitoring the usage of AI algorithms once they are deployed, such as to record potential errors and performance degradation, as well as to perform periodic audits. To improve the explainability of AI algorithms, it is important that AI developers involve clinical end-users from the start of the development process in order to select the best explainability approach for each application and to ensure that the chosen explanations are useful and well accepted in clinical practice. Finally, regulatory entities can play an important role by considering the traceability and explainability of the AI tools as pre-requisites for certification.

3.5. Privacy and security issues

The increasingly widespread development of AI solutions and technology in healthcare, recently highlighted by the COVID-19 pandemic, has shown potential risks for a lack of data privacy, confidentiality and protection for patients and citizens. This could lead to serious consequences (Figure 7), such as the exposure and use of sensitive data which goes against the rights of the citizens or the repurposing of patient data for non-medical gains.

Figure 7 – Main privacy and security risks associated with big data and AI, and some mitigation measures



These issues are firstly linked to informed consent, i.e., the provision of adequate information for the patients for an informed decision such as for sharing personal health data. Informed consent is a crucial and integral part to the patient's experience in healthcare, which was formalised in the Helsinki Declaration and has since grown as the introduction of digital technology has permeated our daily lives (Pickering, 2021). Informed consent is linked to various ethical issues, including protection from harm, respect for autonomy, privacy protection and property rights concerning data and/or tissue (Ploug & Holm, 2016).

However, the introduction of opaque AI algorithms and complicated informed consent forms limits the level of autonomy and the power of shared patient-physician decision making (Vyas et al., 2020). It has become increasingly difficult for patients to understand the decision-making process and the different ways in which their data can be reused, and to know exactly how they can choose to opt out of sharing their data. Issues of informed consent are also especially prominent in big data research, especially digital platform-based health data research, in which a patient may not be fully aware of or fully understand the extent to which their data is shared and reused (McKeown et al., 2021).

An important example of this occurred in 2016, when records of 1.6 million patients in the United Kingdom were transferred – without patients' informed consent – from the Royal Free NHS Foundation Trust to the Google-owned AI company DeepMind, which at the time was working on developing an app to implement new ways of detecting kidney disease (BBC, 2017). In July 2017, the UK Information Commissioner's Office (ICO) ruled that the Royal Free NHS Trust had breached data protection laws; the Information Commissioner office was famously quoted as saying, 'the price of innovation does not need to be the erosion of fundamental privacy rights' (Gerke et al., 2020).

The use of AI in healthcare also entails a risk of data security breaches, in which personal information may be made widely available, infringing on citizens' rights to privacy and putting them at risk for identity theft and other types of cyberattacks. In July 2020, the New York based AI company Cense AI suffered a data breach that exposed highly sensitive data of upwards of 2.5 million patients who had suffered from car accidents, including such detailed information as names, addresses,

diagnostic notes, dates and types of accident, insurance policy numbers and more (HIPPA Journal, 2020). Although eventually secured, this data was briefly accessible to anyone in the world with an internet connection, underlining the very real danger of personal privacy breaches that patients are exposed to.

Another persistent concern is that of data repurposing, which in certain contexts is also referred to as 'function creep' (Koops, 2021). The World Health Organization has warned against the danger of function creep during the COVID-19 pandemic, highlighting a case in Singapore in which the data from the government's COVID-19 tracing applications was also made available for criminal investigations (WHO, 2021). This is a stark example of health-related data being repurposed for non-healthcare related ends, but repurposing can also occur within the healthcare sphere itself. A 2019 report explored in detail the different ways that patient data is repurposed in the European pharmaceutical industry: Data from electronic health records, registry data and data from health systems are used for pharmaceutical drug development, clinical trial design, marketing and cost-effectiveness analyses, and more (Hocking et al., 2019).

In addition to the issues related to data privacy and security, AI tools are especially vulnerable to cyberattacks, the results of which could be anything from burdensome to fatal, depending on the context. In September 2020, a patient died after having to be redirected to another hospital when the Düsseldorf University Hospital suffered a cyberattack that interfered with the hospital's data and rendered the centre's computer system inoperable (Kiener, 2020). Although it was later argued that it could not be proven that the death was directly caused by the cyberattack, because the patient was already suffering a life-threatening condition, this case brought to the forefront the real physical harms that cyberattacks can cause in the healthcare sphere.

In another example of how technological breaches may affect the physical health of patients, in April 2021 the Swedish oncology software company Elekta suffered a healthcare ransomware attack that affected 170 health systems in the United States, delaying cancer treatment care to patients across the country as well as exposing sensitive patient data (Mulcahy, 2021).

Furthermore, research has shown that personal medical devices controlled by AI are also vulnerable to attacks. For example, researchers discovered that AI-powered insulin pumps for diabetes patients could be hacked and remotely controlled from varying distances, and could even be manipulated to flood the patient's body with excessive insulin (Wired, 2019). While this hack has never been carried out in the real world, researchers' development of the AI attack exposed serious vulnerabilities in the AI system's functionality.

These events garnered enough attention to bring to light the question of how algorithmic security – or lack thereof – can affect human survival in a high-stakes context such as healthcare. Focusing on AI tools as part of the larger technological sphere, it is clear that risks of attacks and hacking must be continually monitored.

To address these important issues, there is a need to increase awareness and literacy on privacy and security risks, as well as on informed consent and cybersecurity. Furthermore, regulations and legal frameworks must be extended to address not only privacy but also accountability, and to protect citizens from data breaches and data repurposing. Decentralised, federated approaches to AI should be promoted to leverage the power of big data from clinical centres without the need for unsafe data transfers. Research must be continued and accelerated to improve security in cloud-based systems and to protect AI algorithms against cyberattacks.

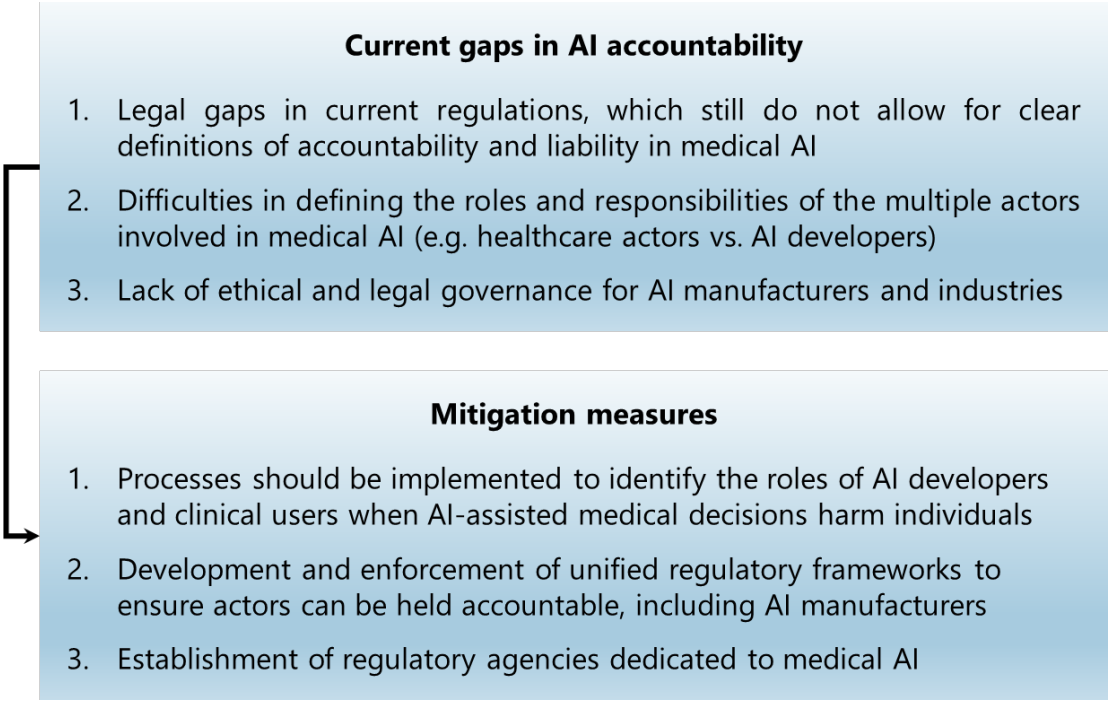
3.6. Gaps in AI accountability

The term 'algorithmic accountability' has garnered increasing importance among researchers and organisations dedicated to addressing the legal impact of the introduction and use of AI algorithms

in different areas of human life. Although the term 'algorithmic accountability' might appear to refer to the task of seeking to hold the algorithm itself accountable, it is actually quite the opposite: It emphasises the fact that algorithms are created through a mixture of machine learning and human design, and that the mistakes or wrongdoings in algorithms come from the humans developing, introducing or using the machines (Kaplan et al., 2018), especially since AI systems themselves cannot be held morally or legally responsible (Raji, 2020).

Accountability is particularly important for medical AI as it will contribute to its acceptability, trustworthiness and future adoption in society and healthcare. For example, clinicians that feel that they are systematically held responsible for all AI-related medical errors – even when the algorithms are designed by other individuals or companies – are unlikely to adopt these emerging AI solutions in their day-to-day practice. Similarly, citizens and patients will lose trust if it appears to them that none of the developers or users of the AI tools can be held accountable for the harm that may be caused. There is a need for new mechanisms and frameworks to ensure adequate accountability in medical AI and to manage reclamations, compensations and sanctions where necessary, as well as to guarantee non-repetition of the acts (WHO, 2021).

Figure 8 – Current limitations in accountability and recommendations to fill in these gaps



Due to the novelty of medical AI and the lack of legal precedence, there is currently a major lack of clarity regarding the definition of responsibilities for AI-related medical errors that could lead to patient harm (Figure 8). The quickly changing and growing field of medical AI poses new challenges for regulators, policymakers and legislators. It pushes current regulations, policies, and laws to adapt their traditional ways of considering responsibility and liability to the new reality of AI-assisted healthcare.

Challenges in applying current law and liability principles to emerging AI applications in medicine include (1) the multi-actor problem in medical AI, which makes it difficult to identify responsibilities among the multiple players involved in the development, implementation and use of medical AI and algorithms (e.g. AI developers, data managers, clinicians, patients, healthcare organisers, etc.); (2) the difficulty in identifying the precise cause of any AI-related medical error, which can be due to the AI algorithm, the data used for training it, or its incorrect use and understanding in clinical

practice; and (3) the multiplicity of governance frameworks and the lack of unified ethical and legal standards in AI industries.

While historically the relationship between the patient and the clinician has stood at the centre of issues concerning medical malpractice and negligence, the introduction of AI tools into healthcare adds a new layer with multiple actors into the patient–physician dynamic (Smith, 2020). These actors may include not only the patient, clinician, healthcare centre, and healthcare system, but also AI developers, researchers, and manufacturers, all of whom are now in some way or another entering into the medical decision-making process. The presence of all these new actors and the lack of clarity – not only on who is responsible for which part of the decision-making process, but also on how the AI tools themselves work – contributes to the complexity of the situation.

While medical professionals are usually under a regulatory responsibility to be able to account for their actions, a requirement that forms an integral part of their professional undertaking, AI developers and technologists generally work under ethical codes (Whitby, 2015). Therefore, for medical professionals the repercussions for not being able to account for their actions and decision-making processes could mean losing their licence to practice medicine; while under the current practice, a lack of accountability for a technologist could mean something much less devastating. Even if an AI manufacturer is found to be responsible for an error, it is often difficult to place blame on one specific person, since so many different developers and researchers work on any given AI system. In addition, the ethical codes and standards of accountability that many private entities use have often been criticised for being vague and difficult to translate into enforceable practice (Raji, 2020).

It is important to note that that the issues of AI accountability and liability in the realm of medicine and healthcare are closely linked to the questions of explainability and transparency. The opaquer an AI algorithm is, the harder it is to find who is accountable for an error involving a patient or a medical decision, and so the burden of responsibility will likely fall more heavily on the clinician who used a non-transparent medical AI tool and is unable to explain their medical decision or the error that occurred (Maliha et al., 2021). This is especially true for assistive AI tools, which are meant to assist the clinician in their decision-making process and may be considered the equivalent of consulting an expert clinical colleague (Harned et al., 2019).

There are avenues to address the current lack of accountability in medical AI. First, processes should be established to identify the roles of AI developers and clinical users when AI-assisted medical decisions harm individuals. There is also a need to establish regulatory agencies dedicated to medical AI. These will develop and enforce regulatory frameworks to ensure specific actors of medical AI can be held accountable, including AI manufacturers.

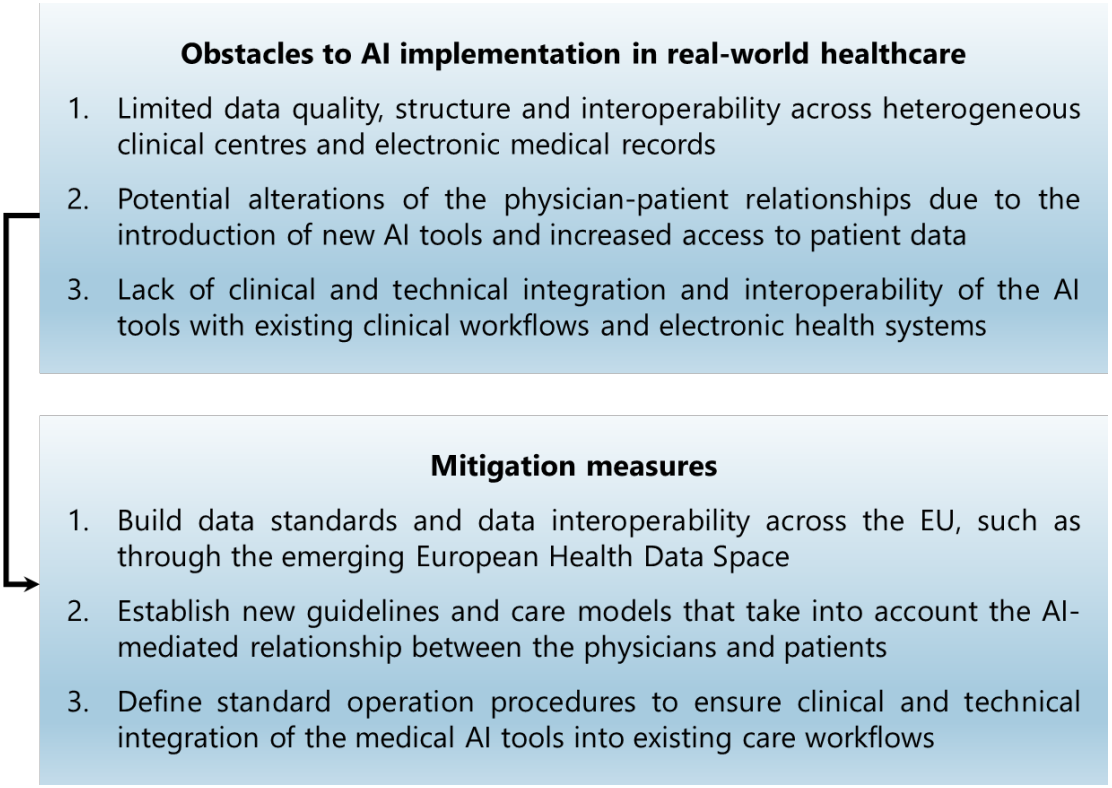
3.7. Obstacles to implementation in real-world healthcare

A large number of medical AI algorithms have been developed and proposed over the last five years, in a wide range of medical applications, as summarised in section 2. However, even when medical AI technologies are well validated and found to be clinically robust and safe, as well as ethically sound and compliant, the road to healthcare implementation, integration and adoption is still laden with specific obstacles in the real world (Shortliffe & Sepúlveda, 2018; Fihn et al., 2019; Nagendran et al., 2020).

Healthcare professionals have traditionally lagged behind other professionals with regards to the adoption of new technologies in their daily activity (Quaglio, 2018). Past experiences in healthcare show that the implementation period is a key stage in the innovation process. In practice, it is not enough to invent and test a new AI technology; other factors which can hinder its implementation in real-world healthcare should also be considered (Arora, 2020), such as (1) the limited data structure and quality in existing electronic health systems, (2) the alteration of the clinician-patient

relationship, as well as (3) the difficulties related to clinical integration and interoperability (Figure 9).

Figure 9 – Obstacles for clinical implementation and integration of new AI tools in real-world healthcare practice, together with potential mitigation measures



First of all, the quality of electronic health data in real-world practice is key to facilitating the implementation of medical AI. However, medical data is notoriously unstructured and noisy, and most existing datasets are not exploitable in AI algorithms. Furthermore, the formats and quality of clinical data vary significantly between clinical centres as well between EU member states (Lehne et al., 2019). Before emerging medical AI tools could be fully implemented and used at large scale, existing data would require significant and costly human revision, quality control, cleaning and re-labelling. To improve data interoperability, the creation of a European Health Data Space was defined as one of the priorities of the European Commission 2019-2025 plan (European Health Data Space). This will promote better re-use of heterogeneous types of health data (electronic health records, genomics data, data from patient registries, etc) across EU countries, including by emerging AI algorithms.

Furthermore, AI technologies are expected to modify the relationship between patients and healthcare professionals in ways that are not yet completely predictable. Certain specialties, particularly those related to image analysis, have already undergone significant transformations due to AI (Gómez-González, 2020 The emergence of patient-centred AI technologies has the potential to transform the historically paternalistic clinician-patient relationship into a join partnership in the decision-making process due to increased transparency and deepened doctor-patient conversations (Aminololama & Lopez, 2019). However, personal and ethical implications of communicating information about AI-derived risks of developing an illness (such as predisposition to cancer or dementia) will need to be elucidated (Fihn et al., 2019; Cohen, 2020). The clinical guidelines and care models will need to be updated to consider the AI-mediated relationships between healthcare workers and patients.

Finally, clinicians and care providers work under established clinical guidelines and technical standards. The introduction of an AI technology into everyday practice will have practical, technical and clinical implications on both clinicians and patients. Secondly, it is not clear that medical AI tools will be systematically interoperable across clinical sites and health systems, and that they will be easily integrated within existing clinical and technical workflows (Meskó & Görög, 2020), without significant modifications to existing clinical practices, care models and even training programmes.

AI manufacturers, in collaboration with healthcare professionals and organisations, will need to establish standard operation procedures for all new AI tools to ensure their clinical interoperability across distinct clinical sites and their integration across heterogeneous electronic healthcare systems. In particular, new AI tools should be developed while ensuring their future integration and communication with already existing technologies, such as genetic sequencing, electronic patient records and e-health consultations (Arora, 2020).

4. Risk assessment methodology

Previous sections of this report have described the main risks that have emerged in recent years concerning the use of AI in healthcare. This calls for a structured approach of risk assessment and management that specifically addresses the technical, clinical and ethical challenges of AI in healthcare and medicine.

4.1. Regulatory frameworks for AI

AI risks can be characterised and classified according to the severity of the harm they may induce, as well as to the probability and frequency of the harm induced. In healthcare, AI risks vary greatly, from infrequent and/or low risks that induce limited and manageable harm to the patients and citizens, to frequent and/or high risks that may cause irreversible damage or harm. For example, an AI algorithm can affect the productivity of the clinicians (e.g. the AI tool fails to accurately delineate the boundaries of the heart in a cardiac image volume, which must be improved manually by the cardiologist), but they can also cause harm to the patient's health and seriously impact the clinical outcomes (e.g. the AI tool fails to diagnose a life-threating condition).

Hence, to minimise the risks of AI and to maximise its benefits in future healthcare, it is important to identify, analyse, understand and monitor the potential risks on a case-by-case basis for each new AI algorithm and application. An important step of the risk assessment procedure should be to devise a methodology for classifying the identified risks into a number of categories representing different levels and types of risk. For each level, a set of tests or regulations must be specified to mitigate and address the AI risks, such that the higher risk classes will require more testing and regulation, while lower risks will result in limited risk mitigation measures. Suitable risk classification of AI according to severity and likelihood will enable manufacturers, care providers and regulators to intervene as much as necessary to ensure the protection of the patients, as well as their rights and values; however, it is also important that these classifications do not –in as much as possible– serve to hamper innovation in healthcare AI.

Currently, the applicable regulations for medical AI tools in the EU are the 2017/745 Medical Devices Regulations (MDR) and the 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR), which were established in 2017. The MDR applies to software as medical devices, including AI-based software, while the IVDR applies to in vitro based diagnostics, including AI-based. These regulations included new approaches for stricter pre-market control, increased clinical investigation requirements, reinforced surveillance across the device's lifecycle, and improved transparency by creating a European database of medical devices. However, many aspects specific to AI are not considered, such as continuous learning of the AI models or the identification of algorithmic biases. In particular, the fact that AI is a highly adaptive technology that continues to learn and adjust over time – as more data becomes available – calls for new approaches to monitor the risks of the AI software.

One of the first proposed for risk assessment in the field of AI came in 2018, when the German Data Ethics Commission proposed to classify risks of general decision-making algorithms according to their criticality, i.e., the system's potential to cause harm (German Data Ethics Commission, 2019). A 'criticality pyramid' comprising five levels of risk/criticality was proposed (1: Zero or negligible potential for harm; 2: Some potential for harm; 3: Regular or significant potential for harm; 4: Serious potential for harm; 5: Untenable potential for harm).

Under this proposal, an adapted testing or regulatory system is recommended depending on the risk level, which could include corrective and oversight mechanisms, specifications regarding the transparency of algorithmic systems and the explainability and comprehensibility of the results, or

rules on the assignment of responsibility and liability within the context of the development and use of algorithmic systems.

In 2021, the European Commission (EC) published a long-awaited proposal for AI regulation and for harmonising the rules that govern AI technologies across Europe, in a manner that addresses safety as well as human rights concerns (European Commission, 2021). In a similar fashion to the 2018 proposal of the German Data Ethics Commission, the draft EU framework provided a definition of AI that is risk-based, together with mandatory requirements for high-risk AI systems. Concretely, the document recommended to classify AI tools according to three main levels of risk: (i) unacceptable risk, (ii) high risk, and (iii) low or minimal risk.

The highest category corresponds to AI tools that contradict EU values and hence should be prohibited. The document (Title II, Article 5) provides some examples of such AI tools, e.g. subliminal manipulation resulting in physical/psychological harm; exploitation of vulnerabilities resulting in physical/psychological harm; social scoring; real-time biometric identification in public spaces (with few exceptions).

The intermediate category, and one of particular interest, corresponds to high-risk AI, which can be permitted only when the tools comply with specific requirements. Such high-risk AI tools (Title III, Chapter 1) comprise safety components of regulated products (including medical devices, but also other products such as toys and machinery), and certain stand-alone AI systems in areas such as operation of critical infrastructure, access to private services as well as employment and workers management. It appears that many medical AI tools, especially those that are autonomous, will be categorised as high-risk. The proposal provides concrete requirements and obligations for adequate risk management in high-risk AI, as listed in Box 1:

Box 1 – Requirements and obligations for high-risk AI tools according to the 2021 EC proposal

Requirements for high-risk AI:

- Use high-quality training, validation and testing data (relevant, representative).
- Draw up technical documentation & set up logging capabilities (traceability & auditability).
- Ensure appropriate degree of transparency and provide users with information on capabilities and limitations of the system & how to use it.
- Ensure human oversight (measures built into the system and/or to be implemented by users).
- Ensure robustness, accuracy and cybersecurity.

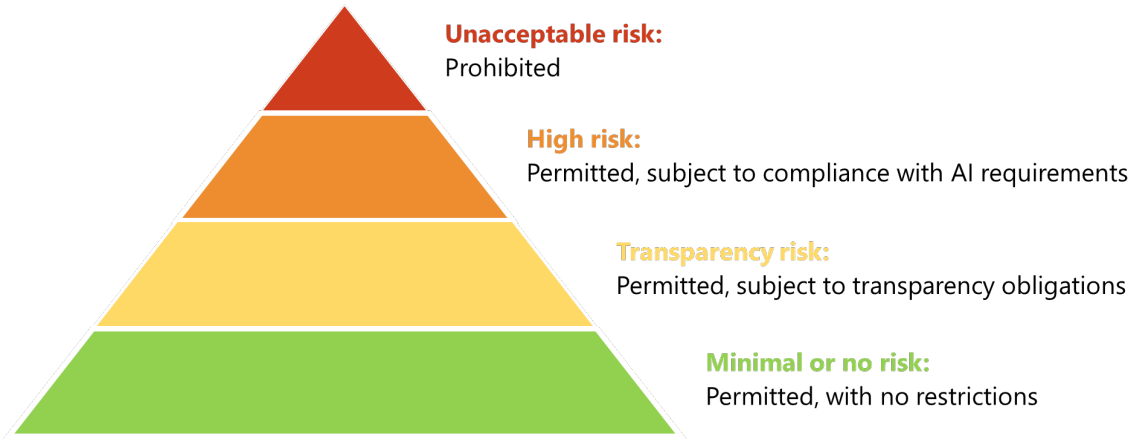
Obligations:

- Establish and implement quality management system in its organisation.
- Draw-up and keep up to date technical documentation.
- Undergo conformity assessment and potentially reassessment of the system (in case of substantial modification).
- Register AI system in EU database.
- Affix CE marking and sign declaration of conformity.
- Conduct post-market monitoring.
- Collaborate with market surveillance authorities.
- Inform the provider or distributor about any serious incident or any malfunctioning.
- Continue to apply existing legal obligations (e.g. under GDPR).

The lowest category refers to AI tools with minimal risk, which have no mandatory obligations but the EC encourages drawing up codes of conduct, as well as voluntary application of requirements for high-risk AI systems or other requirements (Article 69).

In addition to these three categories of risks (unacceptable, high and low), the document (Article 52) discusses an additional category of AI systems, such as those that interact with individuals or expose them to emotional or biometric recognition, for which there is an explicit obligation of transparency. In this case, the individuals must be notified that they are interacting with an AI system (Figure 10).

Figure 10 – AI risk classification according to the 2021 EU proposal on AI legislation



The draft AI regulation does not specifically address AI in healthcare, but it suggests in its current form that AI-driven medical devices will be classified as high-risk, because of the associated safety and privacy concerns. This means future medical AI tools should fulfil all the requirements already established by the Medical Device Regulation, but also those listed in Chapter II of the AI regulation (use of high quality and representative data, technical documentation and traceability, transparency requirement, human oversight, quality management system, conformity assessment, etc).

However, one can argue that not all medical AI tools are systematically high risk. For example, many AI tools have been developed in radiology to accelerate the contouring of organs and lesions on medical images, before quantification and diagnosis (e.g. contouring of the boundaries of the cardiac ventricles or contouring the boundaries of lung tumours). Such AI-powered processing tools are very important and in fact already in use in clinical practice, but they do not necessarily require to be transparent as the clinicians can visually assess the results of the automatic contouring and correct any errors, so the risks are minimal. To continue to promote innovations and investments in medical AI, mechanisms may be needed to discriminate between low- and high-risk AI in healthcare.

With this new regulatory framework, CE marking and regulatory approval in medical AI can take the following form:

- Determine whether the AI tool is classified as high risk under the new AI regulation.
- Ensure AI design, development and quality management systems are in compliance with the AI regulation.
- Undergo conformity assessment procedure to assess and demonstrate compliance.
- Affix the CE marking to the system and sign a declaration of conformity.
- Implement the AI tool in practice or deploy to the market.

It is important to note that the EC proposal for AI regulation is general for all domains of society: it does not take into account the specificities and risks of AI in the healthcare domain. Furthermore, the EC proposal retains of some of the limitations of the MDR and IVDR, such as the lack of

mechanisms to address the dynamic nature of AI technologies. Currently, continuous learning, which is key to medical AI technologies, may be considered as a substantial modification and would require reassessment of the AI technology.

4.2. Risk minimisation through risk self-assessment

For risk identification in AI, several stakeholders have suggested a self-assessment structured approach composed of specified checklists and questions. For example, the independent High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, published an assessment checklist for trustworthy AI called ALTAI. The checklist is structured along seven categories: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability (ALTAI, 2020). In Box 2, some examples of self-assessment questions that were proposed as means to identify potential limitations are provided for reliability, privacy, explainability and fairness:

Box 2 – Examples of self-assessment questions from the ALTAI checklist (ALTAI, 2020)

For reliability:

- Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?
- Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?
- Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?
- Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility?
- Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?
- Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?
- Is your AI system using (online) continual learning?

For data privacy:

- Did you put in place any of the following measures, some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?
  - Data Protection Impact Assessment (DPIA);
  - Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system;
  - Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation);
  - Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system?

For explainability:

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?

For fairness assessment:

- Did you consider diversity and representativeness of end-users and/or subjects in the data?
- Did you test for specific target groups or problematic use cases?
- Did you research and use publicly available technical tools, that are state-of the-art, to improve your understanding of the data, model and performance?
- Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?

The full assessment checklist and questions for all categories can be found online at the Publications Office of the European Union (ALTAI, 2020). It is also available as an online tool for registered users. It is important to note that the list was devised for AI in general and must be tailored to each specific application domain, including healthcare.

To our knowledge, the first self-assessment checklist for AI in healthcare was published by a multi-disciplinary team of researchers from Australia in 2021. Its objective was to help clinicians assess how ready algorithms are for use in routine care and to pinpoint the areas in which further development and finetuning may be necessary before deployment (Scott et al., 2021). This list was put together based on a few narrative reviews on AI in healthcare, which were summarised into a set of assessment questions organised into 10 general questions as listed in Box 3.

Box 3 – Questions from the assessment checklist for medical AI tools, as shown in Scott et al., 2021

- What is the purpose and context of the algorithm?
- How good were the data used to train the algorithm?
- Were there sufficient data to train the algorithm?
- How well does the algorithm perform?
- Is the algorithm transferable to new clinical settings?
- Are the outputs of the algorithm clinically intelligible?
- How will this algorithm fit into and complement current workflows?
- Has use of the algorithm been shown to improve patient care and outcomes?
- Could the algorithm cause patient harm?
- Does use of the algorithm raise ethical, legal or social concerns?

However, this self-assessment list does not contain the same level of detail as the assessment checklist for general AI devised by the AI HLEG. For example, point 10 in Box 3 is rather vague and does enable to pinpoint the exact ethical, legal or social concern (e.g. algorithmic bias). It seems that a combination of both approaches would lead to a detailed and standardised risk assessment checklist for AI in healthcare, generated through consensus and with each category of risk enriched with a detailed set of assessment questions.

This has motivated the recent development of consensus guidelines for trustworthy AI in medicine by a network of EC-funded research projects together with international inter-disciplinary experts. Entitled FUTURE-AI (www.future-ai.eu), these guidelines are organised according to six principles (Fairness, Universality, Traceability, Usability, Robustness, Explainability) and comprise concrete recommendations and a self-assessment checklist to enable AI designers, developers, evaluators and regulators to develop trustworthy and ethical AI solutions in medicine and healthcare (Lekadir et al., 2022). Box 4 lists examples of risk assessment questions included in the FUTURE-AI self-assessment checklist.

Box 4 – Excerpts of risk assessment items from the FUTURE-AI guidelines for trustworthy AI in medicine (version from 27 February 2022)

Fairness:

- Did you design your AI algorithm with a diverse team of stakeholders? Did you collect requirements from a diverse set of end-users?
- Did you define fairness for your specific AI application? Did you ask clinicians about hidden sources of data imbalance?
- Did you thoroughly evaluate the fairness of your AI algorithm? Did you use a suitable dataset and dedicated metrics?

Universality:

- Did you annotate your dataset in an objective, reproducible and standardised way?
- Did you use universal, transparent, comparable, and reproducible criteria and metrics for your model's performance assessment?
- Did you evaluate your model on at least one open-access benchmark dataset that is representative of your model's task and expected real-world data exposure after deployment?

Traceability:

- Did you prepare a complete documentation of the datasets you used? Did you include the relevant metadata?
- Did you keep track, in a structured manner, of the whole pre-processing pipeline of input data? Did you specify input/output, nature, prerequisites and requirements of your pre-processing and data preparation methods?
- Did you record the details of the training process? Did you include a careful description of input predictors?

Usability:

- Did you engage users in the design and development of the AI tool?
- Did you evaluate the usability of your tool after integration in the clinical workflows of the clinical sites?

Robustness:

- Did you train and evaluate your tools with heterogeneous datasets from multiple clinical centres and data protocols?
- Did you evaluate the AI tool under diverse real-world scenarios?
- Did you use any quality control mechanisms to identify potential deviations or artifacts in the input data?

Explainability:

- Did you consult with the clinicians to determine which explainability methods suit them?
- Did you use some quantitative evaluation tests to determine if the explanations are robust and trustworthy? Did you perform some qualitative evaluation tests with clinicians?

The need to further tailor AI risk assessment to specific medical domains have also been stated. For example, in the field radiology, various prominent European and North American radiological associations (American College of Radiology, European Society of Radiology, Radiological Society of North America, Society for Imaging Informatics in Medicine, European Society of Medical Imaging Informatics, Canadian Association of Radiologists, and the American Association of Physicists in Medicine) came together to release a statement on the ethical challenges of using AI in radiology.

They stated that 'the radiology community should start now to develop codes of ethics and practice for AI which promote any use that helps patients and the common good' (Geis et al, 2019).

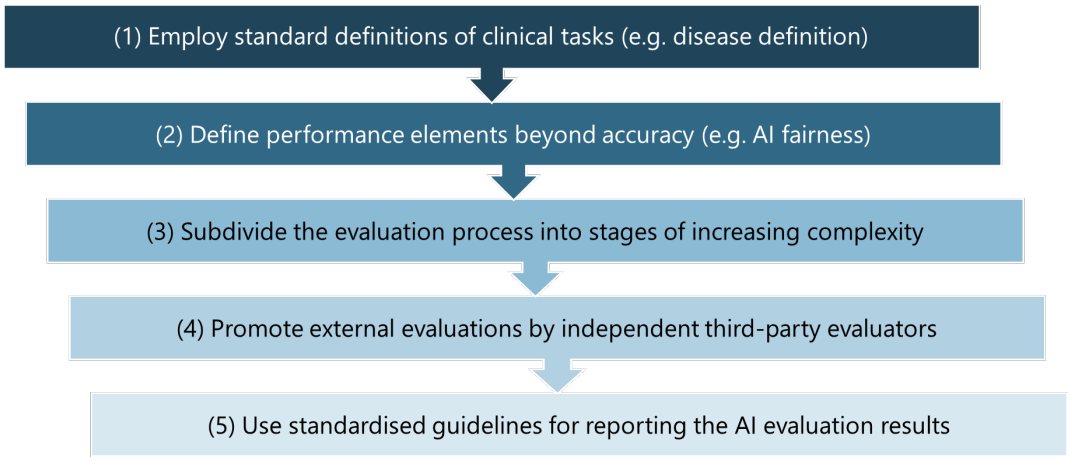
The assessment checklists presented in this section have used different categories of risks, as well as different assessment questions. Standardising, adjusting and validating these approaches through consensus by professional societies and independent groups on a domain-by-domain basis (e.g. radiology vs. surgery) would result in more robust processes for risk identification and management. Furthermore, as more and more healthcare AI algorithms will undergo self-assessment for ethical, legal and technical risks, these checklists should be regularly refined and updated versions will be released for the community taking into account continuous developments in AI methods, processes and regulations.

4.3. Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions

To identify, anticipate and manage risks in medical AI, adequate procedures for evaluating the AI models are of central importance. Thus far, AI evaluation has been achieved mostly by examining model accuracy and robustness in laboratory settings. Other aspects of AI, such as clinical safety and effectiveness, fairness and non-discrimination, transparency and traceability, as well as privacy and security, are more challenging to evaluate in controlled environments and have received less attention in the scientific literature.

Given the existing gaps, the US Food and Drug Administration (FDA) proposed action plan in 2021 to better regulate and advance the agency's oversight of medical AI software, which promoted '*regulatory science efforts to develop methodology for the evaluation and improvement of machine learning algorithms*' (FDA, 2021). In parallel, several research teams have also investigated and proposed new approaches for improved and comprehensive evaluation of medical AI algorithms, especially in North America (Larson et al., 2021; Park et al., 2020), Europe, and Asia (Park & Han, 2018), as well as by international societies such as the International Association of Medical Informatics (Magrabi et al., 2019). In this section, we will summarise their findings into a set of five main recommendations to enable a multi-faceted and comprehensive evaluation of future AI software in healthcare, as outlined in Figure 11.

Figure 11 – Recommendations for improved evaluation of algorithm performance and risks in medical AI



4.3.1. Standardised definition of clinical tasks

To enable objective and comparative evaluation of medical AI solutions, researchers at Stanford University have recently proposed to standardise the definition of the clinical tasks that the AI algorithms are addressing (Larson et al., 2021). In practice, there are many ways to define a clinical task, such as medical diagnostics. As an illustration, the diagnosis and reporting of COVID-19 severity based on chest imaging scans has been proposed using different schemes (Larson et al., 2021), including:

- Two categories: Radiologist's labelling of presence or absence of the disease.
- Four categories proposed by the Radiological Society of North America (RSNA) (Simpson et al., 2020): (1) typical, (2) indeterminate, (3) atypical appearance, and (4) negative for pneumonia.
- Six categories based on the CO-RADS scale (Prokop et al., 2020): (1) negative, (2) low, (3) indeterminate, (4) high, (5) very high, (6) PCR +.
- Various scoring systems of lesion severity in the lungs, such as (i) a 0 to 4 severity rating for each of six lung zones, for a total score of 0 to 24, (ii) a 0 to 5 severity rating for each of five lung lobes, for a total score of 0 to 25, (iii) a 0 to 7 severity rating for each of five lung lobes, for a total score of 35.

Any of these diagnostic systems could be incorporated into an AI-based algorithm, which makes objective assessment of the algorithm's performance and associated risks more difficult. This also limits the ability to directly compare AI-based algorithms that are originally developed for the same clinical task, given the existence of multiple definitions. To date, clinical task definitions have typically been developed with relatively little oversight and coordination. As these clinical tasks will be increasingly performed based on AI algorithms developed by non-clinical developers, it is important that the definitions, which form part of the AI software specifications, should be developed according to accepted consensus-based standard-setting principles and maintained by nonconflicted entities committed to updating the definitions based on new evidence and input from relevant stakeholders. Medical societies, such as the European Society of Cardiology, the European Society of Radiology, or the European Society for Medical Oncology, could play an important role in standardising the definition of the clinical tasks for medical AI in their respective fields. With this approach, the responsibility of the developers will be limited to optimising the performance of the AI algorithms based on widely accepted and utilised reference diagnostic task definitions, which would help ensure widespread acceptance of AI solutions by relevant stakeholders.

4.3.2. Multi-faceted evaluation of performance beyond accuracy

Given the multiple risks and ethical considerations of medical AI, it is now widely accepted that the evaluation of the algorithms must be extended well beyond existing approaches that have mostly focused on model accuracy. While the empirical evaluation of machine learning algorithms remains a matter of on-going debate among researchers, there is a need for the development of specific performance domains for AI in healthcare. Table 2 shows some examples of performance elements recently proposed for AI-based diagnostic algorithms in radiology (Larson et al., 2021). These include classification accuracy, but also reliability, applicability, transparency, monitorability, usability and more (see Table 2).

Table 2 – Examples of performance elements for imaging AI algorithms (from Larson, et. al., 2021)

Accurate	The algorithm should accurately perform all diagnostic tasks for which it is designed.
Reliable	The algorithm should remain accurate in the setting of reasonably expected variation encountered in the clinical environment, including reasonable variations in image quality.
Applicable	The accuracy of the algorithm should be maintained across all makes and models of image modalities and for all patient populations for which it is designed to function.
Deterministic	The algorithm should give the same answer for the same image when used at different times and in different settings.
Non-distractible	The algorithm should be able to recognise the salient information from the image and not change its assessment based on extraneous, non-contributory image data.
Self-aware of limitations	The algorithm should have the means to detect when it is at or beyond the boundaries of its capabilities, whether due to inherent limitations of the model, limitations of its clinical applicability, or limitations imposed by clinical variation such as unexpected patient anatomy or image quality.
Fail-safe	The algorithm should recognise when it has reached an erroneous conclusion and have the means for ensuring that all errors are caught and stopped before they are propagated into the clinical environment
Transparent logic	The user interface should enable the operator to clearly see the linkage between the input and output, including what data were analysed, what alternatives were considered, and why certain possibilities were excluded, to be able to correctly accept or reject the algorithm's conclusion on any given case.
Transparent degree of confidence	The algorithm should share with the user a level of confidence in its assessment for each case. The accuracy of the model's expression of confidence should be validated as well as the accuracy of the model itself.
Able to be monitored	The algorithm should share performance data with users to enable ongoing monitoring of both individual and aggregated cases, quickly highlighting any significant deviations in performance.
Auditable	An independent means should be provided to monitor the algorithm's ongoing performance in a way that guides appropriate intervention. This may include periodic quality control checks similar to those performed by operators on imaging equipment.
Intuitive user interface	The user interface should enable the operator to intuitively how to use the algorithm with as little training as possible and impose the minimum possible cognitive load on the user.

However, it appears that such a list is incomplete, as some important risks of AI in healthcare, such as algorithmic bias and inequality, have not been considered. Among the few works that have directly investigated AI fairness in medicine, it is worth mentioning a recent study that evaluated the state-of-the-art deep neural networks on large public chest X-ray datasets with respect to patient sex, age, race, and insurance type, the latter as a proxy for socioeconomic status (Seyyed-Kalantari et al, 2020). The study concluded that '*models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modification*'. In this work, the authors used the so-called true positive rates (TPR) as a measure of fairness, but other criteria have also been proposed in the literature, such as statistical parity, group fairness, equalised odds and predictive equality (Barocas et al., 2017).

Given the current lack of literacy and trust in AI, clinical usability is another aspect of medical AI that has been recommended for validation with end-users. To enhance clinical acceptance, perceived utility and future adoption, the AI algorithm and its visual interfaces should enable the operator to intuitively know how to use the tool with as little training as possible, to impose the minimum possible cognitive workload on the user, and to enhance clinical efficiency by decreasing decision-making time. During usability tests, questionnaires can be used to gather quantitative and qualitative information on the user's satisfaction with the AI tool (Lewis, 2018). For example, when assessing the usability of an AI-powered algorithm for depression care, the researchers in (Tanguay-Sela et al., 2020) used specific usability questions, as illustrated in Box 5.

Box 5 – Excerpts of a usability questionnaire for assessing an AI technology for depression care (Tanguay-Sela et al., 2020)

- The probabilities produced by the model, overall, were: too optimistic; reasonable; too pessimistic.
- The application interfered with my patient interview: strongly agree; somewhat agree; unsure; somewhat disagree; strongly disagree.
- Based on your overall experience today, how much do you trust the predictive model to help you choose treatments for depression (1 being 'very little' and 5 being 'very much)?
- The model provided us with more rich information to discuss: strongly agree; somewhat agree; unsure; somewhat disagree; strongly disagree.
- Based on your experience today, do you think using the application would cost you significant time (1 being 'cost you significant time' and 5 being 'save you significant time'):
- You would use the application: For all patients with depression; Only for the most severe patients; only for patients where one treatment has failed; only for patients where more than one treatment has failed; not at all; to review patient info.

Other usability elements that could evaluated in a usability questionnaire include: level of understanding of diagnosis by patients and clinicians; level of understanding of treatment options by patients and clinicians; perceived quality of communication between patient and doctor; degree of interpretability of the AI-driven predictions for the clinicians; level of satisfaction with the technology, user interfaces; understanding of technical terminology by clinicians and patients; usefulness of error messages/alerts; overall ease-of-use; impact on clinician's productivity; level of intention-to-use of the system (e.g. only when needed vs. full use), and so on.

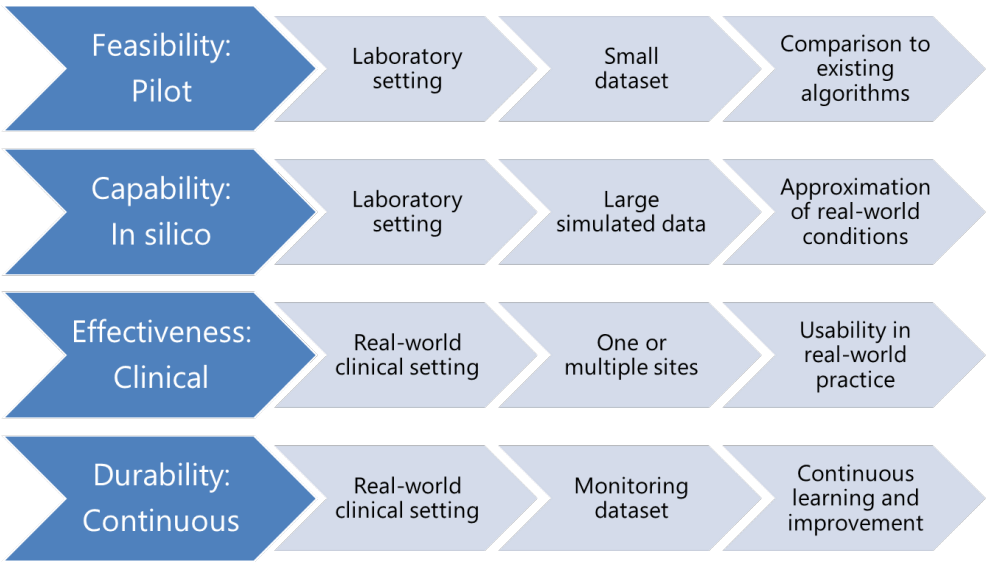
Even if the AI is validated as being accurate, reliable, fair and user-friendly, this may not necessarily lead to patient benefit. Researchers from South Korea suggested assessing impact on patient outcomes to confirm clinical utility and to enable AI technology to be accepted and recommended by clinical experts, academic societies, or independent third-party organisations (Park & Han, 2018). In addition to demonstrating its clinical effectiveness, evaluation of the cost-effectiveness should also be systematically performed, given the huge investments into medical AI with promised efficiencies and cost reductions only being assumed. For example, economic evaluations using decision analytic modelling (Hill et al., 2020) can be used to assess whether additional AI solution costs are justified given the modelled effect, such as on health-related quality of life (e.g. QALY, or quality-adjusted life years). Importantly, the initial investment and operational costs for a given AI infrastructure and service need to be included in the cost-effectiveness analysis (Wolff et al., 2020). Finally, given that AI algorithms continue to learn over time as more data become available, it is important to adapt existing validation frameworks to enable the continuous monitoring of performance throughout the life cycle of the AI tool in the clinical environment.

4.3.3. Subdivision of the evaluation process into discrete phases.

Instead of evaluating medical AI solutions in one single procedure, a few publications have recently recommended implementing a multi-stage approach in which developed algorithms undergo several steps of evaluation of varying goals and increasing complexity. For example, four steps (phase I to phase IV) were proposed for AI validation in the diagnostic imaging field, namely (1) feasibility testing, (2) capability, (3) effectiveness, and (4) durability (Larson et al., 2021) (Figure 12):

- Phase I – Feasibility: The goal is to perform a first/pilot evaluation of the algorithm in the laboratory under ideal conditions, typically on a single small test dataset. This stage will include comparison to existing algorithms that address the same clinical task, or with results obtained directly by expert clinicians. At this stage, the AI algorithms do not need to be fully robust, as the goal is simply to assess feasibility. The resulting findings may be disseminated in a scientific publication, even if the algorithm is not demonstrated for clinical application at this stage.
- Phase II – Capability: In this phase, the goal is to simulate real-world conditions in a laboratory setting and evaluate as well as refine the AI algorithm accordingly to enhance its capabilities. The phase can be also referred to as in-silico validation (Viceconti et al., 2021) (i.e. using computer simulation) or virtual clinical trials (Abadi et al., 2020). In this phase, reliability can be tested by simulating the input data and the clinical conditions under which it may be used. Safety tests will evaluate the algorithm's ability to minimise the risk of harm when deployed and subjected to unanticipated situations, that will be also simulated for testing. Furthermore, this phase should be implemented with end-users, especially clinicians and operators, to evaluate their behaviours and decision making given the simulated conditions and outputs of the AI algorithm.
- Phase III – Effectiveness: At this stage, the validation is moved to the clinical environment to assess real-world performance and to specific clinical sites to perform local validations. The primary objective is to confirm that the real-world performance of the algorithm matches its performance in the test environment. All results and feedback from this stage should be leveraged to update and optimise the AI algorithm, which will be retested in the controlled environment as in previous stages, before another round of local clinical evaluation. This evaluation stage in the clinic may reveal local quality control problems and AI manufacturers should work with local clinical sites to resolve the identified quality issues.
- Phase IV – Durability: At this stage, the manufacturer should put in place a mechanism to enable ongoing performance evaluation and monitoring, with the intent of continuous improvement. They may integrate monitoring or auditing systems within their AI solution to automatically detect, correct, and report errors, and to compile clinical feedback and user feedback. Furthermore, depending on the errors and problems identified over time, the AI algorithms should be updated and improved, such as by using additional training data, and then retested in the controlled environment before they are re-used in the clinic.

Figure 12 – Example of a multi-stage approach for medical AI evaluation



Researchers from IBM Research have proposed an alternative subdivision of the evaluation process by drawing analogies from the drug discovery and testing sectors (Park et al., 2020), as described in Table 3.

Table 3 – Excerpts of subdivided evaluation process for medical AI, based on processes implemented in the drug development sector (Park et al., 2020)

Testing phase of AI algorithm	*Procedures*	*Examples*	*Equivalence in drug discovery*
Phase 1: Technical performance & safety	In silico algorithm performance optimisation Usability tests	Determination of thresholds to balance sensitivity and specificity for a particular clinical use case, scenario-based testing to assess cognitive overload	Determine optimal dose Identify potential toxicities
Phase 2: Efficacy & side effects	Controlled algorithm performance/efficacy evaluation by intended users in medical setting Interface design Quality improvement	Retraining and reassessing model performance with larger real-world data sets, measurement of the efficiency of information delivery and workflow integration with representative users, pilot study of predictive algorithm in a clinical setting	Early efficacy tests Adverse event identification
Phase 3: Therapeutic efficacy	Clinical trial Adverse events identification	Randomised trial to test whether delivery AI-based decision support affects clinical outcomes and/or results in user over-trust	Clinical trial Adverse event identification
Phase 4: Safety & effectiveness	Post-deployment surveillance	Measurement of algorithmic performance drift	Post-marketing surveillance

While there are overlaps between the two subdivisions of the medical evaluation process presented in this section (Figure 12 & Table 2 – Examples of performance elements for imaging AI algorithms (from Larson, et. al., 2021)). The first subdivision (Figure 12) is focused on separating the environments

and populations in which the algorithm is tested (small datasets to demonstrate feasibility, simulated environments to test robustness to contextual changes, clinical setting to demonstrate real-world applicability). The second approach (Table 2) does not necessarily separate the testing environments (e.g. medical settings are used in both phases 2 and 3) but each step is more focused on a particular risk and clinical aspect such as on safety, effectiveness, usability and efficacy.

In both multi-stage evaluation approaches, each of these phases is dependent upon the successful completion of the previous step, which reduces costs. For example, algorithms that do not perform well in a controlled environment are almost certain to not perform well in the real world. While they require to be further developed and adopted by the relevant stakeholders, these multi-stage and multi-faceted evaluation studies are promising as they take into consideration the complexity of AI-guided healthcare delivery, which is compounded by user- and context-dependent applications.

4.3.1. Promotion of external evaluations by third-party evaluators

Evaluating the performance of an AI model with similar datasets than those used to develop and train the model is called internal validation. In the early days of medical AI, this was the most reported approach for algorithm validation as it is easy to implement. However, internal validation – even by developers and manufacturers with a culture of quality and good practices of excellence in medical AI – is likely to be inherently biased and to overestimated performance, while it is limited in its ability to identify all risks associated with changes in the data or clinical environment. A 2019 study reviewed more than 500 research papers in the field of radiology AI and found that only 6% of the AI algorithms reported underwent an external evaluation (Kim et al., 2019). Hence, in recent years many researchers and opinion leaders have recommended promoting the external evaluation of AI algorithms in healthcare (Park & Han, 2018; Larson et al., 2021).

External validation refers to the use of completely separate, external datasets for evaluating AI tools. The external datasets should strongly represent the variability in the population and the usage of the AI solution. Such data will ideally come from different clinical sites and geographical locations to evaluate the generalisability of the given AI algorithm outside of the controlled environment in which it was built. With this approach, it will be possible, for example, to evaluate the AI algorithm when the technical parameters of the data acquisition vary (e.g. differences in imaging scanners and protocols between hospitals). Furthermore, many researchers have recommended the use of common reference datasets, acquired from representative real-world populations, for external evaluation and benchmarking of AI models. These reference datasets can be directly compared to similar algorithms that have been previously evaluated with the same reference dataset. For example, in 2010 the National Cancer Institute in the United States set up the Cancer Imaging Archive (www.cancerimagingarchive.net), which now comprises a wide range of cancer imaging collections from all cancer types, that are extensively and routinely used for external validation and comparison of AI algorithms.

Several research projects have recently been funded by the European Commission to build European repositories of reference cancer imaging datasets, such as the EuCanImage project (https://eucanimage.eu). Furthermore, external validation should ideally be carried out by using third-party evaluators to ensure an objective and exhaustive evaluation of the AI algorithm is performed according to the performance criteria outlined in the previous section, such as accuracy, reliability, fairness and usability. Such third-party evaluators could include clinical research organisations, research laboratories, or independent institutions that develop and maintain reference standard data sets. Such testing organisations would be specialised to enable the highest standards, quality and objectivity in the evaluation and monitoring of AI solutions in healthcare, resulting in reduced undetected risks and increased trust in medical AI for real-world practice. It is worth noting that DIGITAL EUROPE is currently preparing new research initiatives to develop Testing and Experimentation Facilities (TEF) in Europe, which -once established- will greatly facilitate external validation of medical AI tools, especially for companies.

4.3.2. Standardised and comprehensive reporting of the AI evaluation procedure and results

To further enhance trust and usability of the AI tools, transparent documentation and reporting of the validation process is essential. This type of reporting will facilitate the critical appraisal process for developers, researchers, and other stakeholders; in addition, it should help replicate the AI algorithm and results, if necessary. Before the widespread use of AI, researchers had already identified the need for standardised and comprehensive reporting guidelines for predictive models used in healthcare, among which is TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (Collins et al., 2015). The TRIPOD statement was first published in 2015 and shortly afterwards adopted at large in the biomedical community. TRIPOD provides guidance on how to clearly report the development of a predictive model in order to assess its potential bias and usefulness. (Collins et al, 2015) Concretely, and as illustrated in Box 5, the TRIPOD statement includes a checklist of 22 items deemed essential for transparent reporting of a prediction model study.

Box 5 – Essential items to be included when reporting a prediction model, according to TRIPOD

- Title, abstract, background, and objectives.
- Methods: Source of data, participants, predictors, sample size, missing data, type of prediction model and other model-building procedures, etc.
- Results: Participants (number and characteristics), performance measures, confidence intervals, model updating., etc.
- Discussion: Limitations (e.g. non-representative sample, missing data), interpretation (incl. comparison to similar studies), implications (e.g. potential clinical use).
- Other information: Supplementary information, funding.

Although TRIPOD primarily aims to improve reporting, it also facilitates more comprehensive understanding and analysis of prediction models, ensuring that they can be further studied and used to guide the provision of healthcare, thus enhancing reproducible research, trust and clinical translation. While many aspects of the TRIPOD statement are inherently applicable to prediction model studies using machine learning methods, its uptake by AI communities has not been high. Possible reasons for the low level of uptake include subtle differences in terminology or a perceived lack of relevance because TRIPOD – at least in its original definition – focused on regression-based prediction model approaches (and not machine-learning based ones). In response to more AI-specific reporting guidelines, an extension of TRIPOD devoted to health prediction models that use machine learning techniques is currently being developed under the name of TRIPOD-AI (Collins et al, 2021)[1].

Another example of reporting and validation guidelines is the work carried out by the CONSORT consortium (Consolidated Standards of Reporting Trials), which has extended their 2010 reporting guidelines to include AI-specific aspects with their CONSORT-AI statement. While the original guidelines recommended including elements such as title, trial design, participants, interventions, outcomes and sample size, the extended CONSORT-AI statement proposes that researchers 'provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human–AI interaction and provision of an analysis of error cases' (Liu et al, 2020). As shown in Box 6 (Liu et al, 2020), the CONSORT-AI extension enumerates new AI-specific items to be used in

[1] TRIPOD. www.tripod-statement.org

the reporting process, in addition to those included in the original CONSORT guidelines published in 2010.

Box 6 – Reporting elements for medical AI in clinical trials, according to the CONSORT-AI guidelines

- Indication that the intervention involves AI in the title and abstract and specify the type of model.
- Intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, public).
- Description of how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.
- Version of the AI algorithm that was used.
- Description of the input data that were acquired and selected for the AI intervention.
- Description of any human–AI interaction in the handling of the input data, and the level of expertise required from users.
- The output of the AI intervention.
- Explanation on how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.
- Results of any analysis of performance errors and how errors were identified, where applicable.
- Information on how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.

Researchers at Stanford University proposed a new set of standards for reporting AI solutions in healthcare, entitled MINMAR (MINimum Information for Medical AI Reporting) (Hernandez-Boussard et al., 2020). The MINMAR standards describe the minimum information necessary to understand intended predictions, target populations, model architecture, evaluation processes, and hidden biases. The MINMAR guidelines are specifically designed for medical AI and comprise reporting elements in four main categories, as shown in Table 4.

Table 4 – Reporting elements from the MINMAR reporting guidelines

Element	Description
1. Population & setting	
Population	Population from which study sample was drawn
Study setting	The setting in which the study was conducted.
Data source	The source from which data were collected
Cohort selection	Exclusion/inclusion criteria
2. Patient demographic characteristics	
Age	Age of patients included in the study
Sex	Sex breakdown of study cohort
Race/ethnicity	Race/ethnicity breakdown of patients included in the study
Socioeconomic status	A measure or proxy measure of the socioeconomic status of patients included in the study

3. Model properties	
Model task	Classification or prediction
Model architecture	Algorithm type: Machine learning, deep learning, etc.
Data splitting	How data were split for training, testing, and validation
Gold standard	Labelled data used to train and test the model
Features	List of variables used/selected in the AI model
Missingness	How missingness was addressed: reported, imputed, or corrected
Optimisation	Model or parameter tuning applied
Internal model validation	Study internal validation
External model validation	External validation using data from another setting
Transparency	How code and data are shared with the community

Such a reporting model for medical AI evaluation will promote transparency, thoroughness, and trust, by including all the key information from the AI evaluation studies in a single detailed document, as well as by assisting publishing editors, AI developers, clinicians and researchers in understanding, interpreting and critically appraising the quality of the AI study design, validation and results.

5. Policy options

This section describes seven policy options suggested to better develop, evaluate, deploy and exploit technically, clinically and ethically sound AI solutions in future healthcare (Figure 13).

Figure 13 – Summary of policy options suggested in this report



5.1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements

As described in Section 4.1, current medical AI devices are regulated by the MDR and IVDR regulations, which were introduced in 2017. Furthermore, in 2021 the European Commission (EC) proposed a new regulation for AI which provides new requirements and obligations for high-risk applications, including medical AI technologies, such as to establish and implement quality management systems in organisations, undergo conformity assessment and potentially reassessment of AI systems (in the event of substantial modification), as well as conduct post-market monitoring.

While the new proposal has been elaborated for AI technologies in general, the new framework considers medical AI tools as high risk, requiring them to undergo increased scrutinisation. However, the requirements are presented in a generic fashion, while – as seen in this report – AI in healthcare is faced with specific and high-stake technical, clinical and socio-ethical challenges and risks.

It is thus important that regulatory frameworks and codes of practice are extended and put into practice for medical AI (as described in sections 4.2 and 4.3). The need for updating the regulatory approvals of AI-driven medical devices has been voiced worldwide, such as in the United States (Harvey & Gowda, 2020; Allen, 2019), Japan (Chinzei et al., 2018; Ota et al., 2020) and China (Roberts

et al., 2020). Particularly, in 2021 the U.S. Food & Drug Administration (FDA) published the Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan (FDA, 2021), which calls for tailored regulations for medical AI, good machine learning practices, and patient-centred approaches.

For tailoring existing frameworks and AI practices to the medical field, multi-faceted risk assessment (section 4.2) should be an integral part of the medical AI development and certification process. Furthermore, risk assessment must be domain-specific, as the clinical, social and ethical risks and constraints differ between, for example, radiology, surgery, genomics, mental health, child health, and home care.

The validation of medical AI technologies should be harmonised and strengthened to assess and identify multi-faceted risks and limitations by evaluating not only model accuracy and robustness but also algorithmic fairness, clinical safety, clinical acceptance, transparency and traceability.

An important proposal (highlighted in section 4.3) for improved medical AI validation and certification is the introduction and generalisation of third-party external validation by independent entities that will be specialised in this process. This will allow for a more objective and expert validation of medical AI tools in a manner that systematically takes into account variability in real-world clinical practices and socio-ethical contexts.

5.2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms

For the future acceptability and implementation of medical AI tools in the real world, many stakeholders beyond AI developers – such as clinicians, patients, social scientists, healthcare managers and AI regulators – will play an important role. Hence, new approaches are needed to promote inclusive, multi-stakeholder engagement in medical AI and ensure the AI tools are designed, validated and implemented in full alignment with the diversity of real-world needs and contexts.

Hence, future AI algorithms should be developed by AI manufacturers based on co-creation (Leone et al., 2021), i.e. through strong and continuous collaborations between the AI developers and the clinical end-users, as well as with other relevance experts such as biomedical ethicists. These collaborations should be present at all stages, from the design and development of the AI solution to its validation and deployment (Filice & Ratwani, 2020).

Integrating human- and user-centred approaches throughout the whole AI development process will enable to design AI algorithms that better reflect the needs and cultures of healthcare workers, but also to identify and address potential risks at an early stage. This will shift the focus towards optimising the clinical performance of the end-users and the health benefits for the citizens, while considering existing social, ethical and legal requirements.

Through strong user engagement, future implementations of medical AI algorithms will take into close consideration the expected interactions between the end-users and the algorithms (otherwise referred to as human-computer interaction) (Xu, 2019). Visual interfaces should be carefully designed based on requirements from the clinical end-users to enable human-centred and clinically meaningful displays of explanations for the machine learning model predictions in healthcare (Barda et al., 2020). This will allow human errors to be reduced and will improve explainability and acceptance of the AI-driven predictions and decisions.

Finally, multi-stakeholder engagement and co-creation will address specific social issues related to equity, equality and fairness, which are application-specific issues that require an understanding of the clinical tasks, possible confounding factors, and relevant group differences; hence continuous

collaboration between the domain experts, healthcare professionals, social scientists, and real-world community members, especially from underrepresented groups, is key.

5.3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI

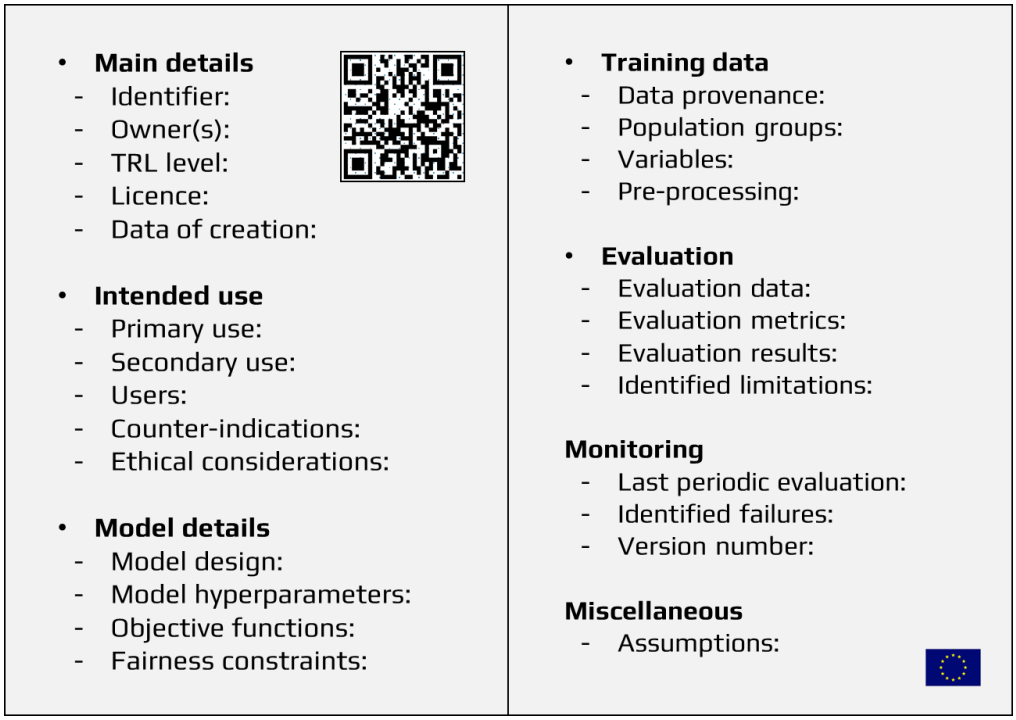
New approaches and mechanisms are needed to enhance the transparency of AI algorithms throughout their lifecycle. In order to be able to understand the details of what has occurred when something goes wrong in the clinical implementation of medical AI, transparency is essential, including but not limited to documenting the whole AI development process; this type of documentation and transparency helps eliminate potential ambiguities and lack of accountability (Felzmann et al, 2020).

One option is for regulatory bodies for medical AI to introduce an 'AI passport' for standardised description and traceability of medical AI tools (see illustration in Figure 14). Such a passport should describe and monitor key information about the AI technology, covering at least five categories of information:

1. Model related information (e.g. model owners, developers and reviewers, intended clinical uses, applicable licences(s), algorithmic details, hyper-parameters, key assumptions and requirements).
2. Data related information (training vs. testing data, data types e.g. imaging, real vs. simulated datasets, data origins).
3. Evaluation related information (model accuracy, robustness, biases, limitations and extreme cases).
4. Usage related information (e.g. statistical distributions, (dis)agreements with clinicians, identified failures, memory usage, etc.).
5. Maintenance related information (last updates, software versioning, last periodic evaluation, dates, etc.).

The AI passport should be standardised to enable consistent traceability across countries and healthcare organisations.

Figure 14 – Example of a possible AI passport that can be used to improve traceability and transparency in medical AI, by documenting all key details about the AI tools, their intended use, model and data details, evaluation results, and information from continuous monitoring and auditing



Furthermore, medical AI is a highly dynamic technology with new data, equipment and users regularly introduced into its workflows. It is therefore clear that the concept of traceability must go beyond the mere documentation of the development process or the phase of testing the AI model; instead, it should also comprise the process of monitoring and maintaining the AI model or system in the real world by continually tracking how it functions after deployment in clinical practice and identifying potential errors or changes in performance (Lekadir et al, 2022).

Hence, it is important that the algorithms are developed together with accompanying live interfaces that will be intended for continuous surveillance and auditing of the AI tools after their deployment in their respective clinical environment. Such monitoring tool should include user-friendly capabilities for quality control and detection of errors and extreme cases, a human-in-the-loop mechanism to enable for human oversight and feedback, a system of alerts to inform the clinicians of suspected deviations from previous states or performance degradation (e.g. when new equipment or protocol is introduced), as well as a periodic evaluation system that can be configured to indicate reference test datasets, as well as periodicity of the evaluations (e.g. monthly vs. quarterly).

5.4. Develop frameworks to better define accountability and monitor responsibilities in medical AI

Accountability continues to be a pressing issue in the field of AI, especially in the high-stake areas of medical AI. It is an especially important issue when considering situations in which an AI-based healthcare tool deployed in real clinical settings fails, produces errors, or results in unexpected side effects (Geis et al, 2019). Frameworks and mechanisms are needed to adequately assign responsibility to all actors in the AI workflow in medical practice, including the manufacturers, thus providing incentives for applying all measures and best practices to minimise errors and harm to

the patient. Such expectations are already an integral part in the development, evaluation and commercialisation of medicines, vaccines and medical equipment, and need to be extended to future medical AI products.

Above all, unified legal frameworks are needed to define responsibility and liability and enforce relevant consequences in medical AI across Europe and beyond. Of the existing regulations, the GDPR offers a two-pronged approach to algorithmic accountability – approaching the issue from the perspective of individual rights on the one hand and systemic regulatory frameworks on the other (Kaminski & Malgieri, 2019). In particular, the GDPR establishes transparency as a key principle for data processing and links it with lawfulness (Art. 5 para 1(a) GDPR) which both are important parts of the principle of accountability (Art. 5 para 2 GDPR).

However, while the GDPR is highly variable in terms of outlining the rights to data privacy as well as to explanation, some researchers in the field have stressed that it is not in and of itself sufficient in terms of outlining algorithmic accountability in medical AI (Barocas, 2019). There is a legal gap for medical AI accountability that remains to be addressed; in the face of this challenge, expert leaders in the field have recommended the establishment of a singular new regulatory body for AI (Tuut, 2017; Koene et al., 2019).

It is expected that in 2022 the EC will propose EU-wide measures adapting existing liability frameworks to the challenges of AI in order to ensure that victims who suffer damages to their life, health or property from an AI technology have access to the same compensation as victims of other technologies (Communication to EU Parliament, 2021). This may include a revision of the Product Liability Directive (Council Directive, 1985) and may require sectorial adjustments such as for AI in healthcare.

One important way of increasing accountability of AI tools in healthcare is through periodic audits and risk assessments, which can be used to evaluate how much regulatory oversight a certain AI tool might need (Kaminski & Malgieri, 2019; Reisman et al., 2018). To this end, the assessments must be conducted through the whole AI pipeline, from data collection, to development, to pre-clinical stages, to deployment, but also when the tools are in use. Future AI solutions should maintain an archive of AI-based decisions and have a mechanism for continuous monitorability and traceability over time as described in the previous section. Audits to assess fairness, transparency, accuracy, and safety could be used to hold AI decision-making processes to the same standard as human processes (Caplan et al., 2018). While some companies and agencies lean heavily on internal auditing processes, numerous researchers as well as civil rights organisations call for these audits to be carried out externally by independent auditing organisations.

5.5. Introduce education programmes to enhance the skills of healthcare professionals and the literacy of the general public

To increase adoption and minimise error, future medical professionals need to be adequately trained in this new technology, including its advantages to improve care, quality, and access to healthcare, as well as its limitations and risks (Paranjape et al., 2019). Hence, it is time to update educational programmes in medicine and increase their interdisciplinarity, with dedicated lectures and practical sessions that seamlessly integrate the implications of medical AI in future clinical practice (McCoy et al., 2020; Rampton et al., 2020).

Furthermore, there is an urgent need to increase the AI literacy of the general public to empower citizens and patients, who will better seize the benefits of emerging medical AI tools, while minimising the potential risk of misuse of the AI tools, especially during remote monitoring and care management. Some countries have already invested in providing free AI public literacy courses, such as Finland's 'Elements of AI' course run by the University of Helsinki (www.elementsofai.com).

5.6. Promote further research on clinical, ethical and technical robustness in medical AI

Despite major advances in recent years in AI and machine learning, as well as in their applications to medicine and healthcare, the multitude of risks discussed in this report call for further research and development to realise the full promise of medical AI, while addressing the existing clinical, socio-ethical and technical limitations. Examples of areas for future research include explainability and interpretability, bias estimation and mitigation, as well as secure and privacy-preserving AI.

Explainable AI is a research area that is investigating a new generation of AI algorithms that can be understood by humans, such as by clinicians and patients in medical AI. It has attracted a lot of interest in recent years and various approaches are being developed and tested. However, explainable AI in healthcare remains very challenging due to the complexity and variability of the biomedical and clinical data, and existing methods are yet to find their way to clinical practice. To improve their potential, it is important to assess and ensure that explainability methods produce explanations that are clinically meaningful and accepted by the end-users. There is a need for interdisciplinary approaches during AI developments that start by examining the needs of the clinicians and understanding the types of explanations (visual vs. quantitative methods) that better suit their needs and specific clinical task.

To explicitly mitigate the presence of unwanted bias in the data, methods have already been investigated (Li & Vasconcelos, 2019; Zhang et al., 2018) and some open-source toolkits have already been published, such as those by IBM (AI Fairness 360) and Microsoft (Fairlearn (Bird et al., 2020)). However, the detection of biases, in particular implicit and hidden biases, remains to a great extent an open problem. Qualitative biases such as cognitive biases of clinicians generating, interpreting or annotating the data, require multidisciplinary research and increased diversity in AI development, healthcare, and policy teams to mitigate bias and strengthen the fairness of medical AI algorithms.

There is also need for more research to develop adaptation methods that will ensure a high level of generalisability of future AI tools across population groups, clinical centres and geographical locations. In addition, it is important to develop new validation platforms that can robustly assess AI algorithms for accuracy but also for fairness with respect to sex/gender, age, ethnicity and race, socioeconomic status and other sociodemographic categories.

Furthermore, future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions (Kompa et al., 2021). Ideally, the clinician should receive alerts/warnings when the uncertainty for certain predictions is high. In future settings, the AI system could provide information on the cause of the high uncertainty (e.g. low-quality image scans, insufficient evidence in the data), and even advise the clinicians on the course of action needed to improve the AI predictions (e.g. inclusion of additional lab tests and predictors, re-scanning of the patient).

Finally, current cyberattacks on medical AI technologies remain difficult to detect, as the actual tools themselves may continue to function properly, but the conclusions that the AI system will confidently provide will be erroneous. Further research is needed to develop, validate and deploy medical AI tools that are able to protect themselves against privacy as well as security risks. This will result in a new generation of AI algorithms which can be robustly deployed and used in their real-world environment with maximal resilience and confidence.

5.7. Implement a strategy for reducing the European divide in medical AI

While the EU has made significant investments in AI in recent years, inequalities persist between different European countries when it comes to advancements in the field of AI (Caradaica, 2020). The AI divide – especially between the Western and Eastern regions of the continent – can be explained by structural differences in research programmes and technological capacities, as well as by the varying levels of investments from the public and private sectors (Quaglio, et al., 2020B).

The disparities in AI development and implementation between EU countries are particularly marked in medical AI, since developments and innovations in this field are highly dependent on access to large databases of well-curated biomedical data as well as to technological capacities. At the same time, these AI disparities may exacerbate the existing health inequities and disparities that exist across the EU; for example, studies have shown that there is a gap between Eastern and Western Europe in life expectancy, maternal mortality, and other population health indicators (Forster, 2018; The World Bank, 2019).

In this context, the EU Member States, in particular those of Eastern Europe, could develop specific programmes to support AI in health. These should include concrete actions to boost the technological, research and industrial capacities of emerging EU countries in the field of AI for healthcare. In particular, infrastructure projects should be established by Member States that have limited research infrastructures and data availability. This would build and enhance much-needed capacities in biomedical and health data sharing, storage, curation and security across the entire EU (ECRIN, 2019). Other programmes should be established to increase the technological, clinical and industrial capacities of several European countries for the development, testing and deployment of novel AI tools in medicine and healthcare, including high-performance computing, open cloud services, clinical testing facilities and pre-commercial procurement.

The European Commission could implement specific coordination and support programmes of activities implemented in this sector by different Member States, thereby supporting the implementation of common guidelines and approaches. Such coordination should ensure the development of an inclusive European Health Data Space (EHDS), which takes into close consideration national and regional challenges across Europe (Marschang, 2021). Similarly, existing education-focused programmes such as the Marie-Curie training networks could be strengthened to enhance the training capacities and human capital in medical AI specifically in emerging EU countries.

Lastly, the disparities that exist in medical AI between different European countries – and especially between Eastern and Western Europe – also reflect the broader social, economic, and health inequities across the different regions of Europe. The issue of reducing the European divide in medical AI is one that requires an approach that goes beyond focusing solely on the fields of medicine and/or the fields of AI and instead involves policy actions that will tackle the larger issues of systemic inequality in European society.

References

Abadi, E., Segars, W.P., Tsui, B.M., Kinahan, P.E., Bottenus, N., Frangi, A.F., Maidment, A., Lo, J. and Samei, E., 2020. 'Virtual clinical trials in medical imaging: a review', Journal of Medical Imaging, 7(4), p.042805.

Abdi J, Al-Hindawi A, Ng T, Vizcaychipi MP. 'Scoping review on the use of socially assistive robot technology in elderly care', BMJ Open. 2018;8(2):e018815.

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N. and Folk, J.C.. 'Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices', NPJ digital medicine, 1(1), pp.1-8., 2018

Abràmoff, M.D., Tobey, D. and Char, D.S. 'Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process,' American journal of ophthalmology, 214, pp.134-142, 2020

Adamson, A.S. and Smith, A., 'Machine learning and health care disparities in dermatology' JAMA dermatology, 154(11), pp.1247-1248, 2018.

Adedinsewo D, Carter RE, Attia Z, Johnson P, Kashou AH, Dugan JL, et al. 'Artificial Intelligence-Enabled ECG Algorithm to Identify Patients with Left Ventricular Systolic Dysfunction Presenting to the Emergency Department with Dyspnea'. Circ Arrhythmia Electrophysiol.;13(8), 2020

Adhikari L, Ozrazgat-Baslanti T, Ruppert M, Madushani RWMA, Paliwal S, Hashemighouchani H, et al. 'Improved predictive models for acute kidney injury with IDEA: Intraoperative data embedded analytics' PLoS One. 2019;14(4).

Ahn J, Connell A, Simonetto D, Hughes C and Shah VH. 'Application of Artificial Intelligence for the Diagnosis and Treatment of Liver Diseases,' Hepatology. 2021;73(6):2546-2563.

Alder, S. 'AI Company Exposed 2.5 Million Patient Records Over the Internet', HIPPA Journal. 21 August 2020.

Allen M, Pearn K, Monks T, Bray BD, Everson R, Salmon A, James M, Stein K. 'Can clinical audits be enhanced by pathway simulation and machine learning? An example from the acute stroke pathway', BMJ Open. 2019;9(9):e028296.

Allen, B., 'The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices', Journal of the American College of Radiology, 16(2), pp.208-210, 2019.

Almirall, D., Nahum-Shani, I., Sherwood, N.E. and Murphy, S.A., 'Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research', Translational behavioral medicine, 2014, 4(3), pp.260-274.

Alsharqi M, Woodward WJ, Mumith JA, Markham DC, Upton R, Leeson P. 'Artificial intelligence and echocardiography', Echo Res Pract. 2018;5(4):115–25.

Aminololama-Shakeri S, Lopez E. 'The Doctor-Patient Relationship With Artificial Intelligence', American Journal of Roentgenology. 2019;202(2)

Andrew D Selbst, Julia Powles., 'Meaningful information and the right to explanation', International Data Privacy Law, Volume 7, Issue 4, Pages 233–242, 2017

Angus DC. 'Randomized clinical trials of artificial intelligence', JAMA. 323(11):1043-1045, 2020.

Arora A. 'Conceptualising Artificial Intelligence as a Digital Healthcare Innovation: An Introductory Review', Med Devices (Auckl). 3:223-230. doi: 10.2147/MDER.S262590, 2020.

Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. 'Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs', Circ Arrhythmia Electrophysiol. 12(9), 2019

Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. 'An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction', Lancet. 394(10201):861–7, 2019.

Azzi, S.; Gagnon, S.; Ramirez, A.; Richards, G. 'Healthcare Applications of Artificial Intelligence and Analytics: A Review and Proposed Framework', Appl. Sci., 10, 6553. https://doi.org/10.3390/app10186553, 2020.

Baetan, R., Spasova, S., Vanhercke, B., Coster, S., Inequalities in access to healthcare: A study of national policies, European Commission, 2018.

Bandivadekar, S.S., 'Online Pharmacies: Global Threats and Regulations', AAYAM: AKGIM Journal of Management, 10(1), pp.36-42, 2020.

Barda, A.J., Horvat, C.M. and Hochheiser, H., 'A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare.', BMC medical informatics and decision making, 20(1), pp.1-16, 2020.

Barocas, S. Legal and Policy Implications of Model Interpretability. This Week in Machine Learning and AI (TWiMLAI), January 2019. https://twimlai.com/twiml-talk-219-legal-and-policy-implications-of-model-interpretability-with-solon-barocas/.

Barocas, S., Hardt, M. and Narayanan, A., 'Fairness in machine learning', Nips tutorial, 1, p.2, 2017.

BBC, 'Google DeepMind NHS app test broke UK privacy law', BBC News, 3 July 2017.

Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-715.

Berlyand Y, Raja AS, Dorner SC, et al. How artificial intelligence could transform emergency department operations. Am J Emerg Med. 2018;36(8):1515-1517.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. and Walker, K., 'Fairlearn: A toolkit for assessing and improving fairness in AI,' Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.

Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, R Van Meter A, De Choudhury M, Kane JM. 'Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook', NPJ Schizophr.;5(1):17, 2019.

Boniolo F, Dorigatti E, Ohnmacht AJ, Saur D, Schubert B, Menden MP. 'Artificial intelligence in early drug discovery enabling precision medicine', Expert Opin Drug Discov:1-17, 2021.

Campello, V. et al. 'Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge.' Medical Image Computing and Computer Assisted Intervention, 2020.

The Cancer Imaging Archive. www.cancerimagingarchive.net, accessed November 2021.

Caplan, R., Donovan, J., Hanson, L. and Matthews, J., Algorithmic accountability: A primer. Data & Society, 18, 2018.

Caradaica, M., 'Artificial Intelligence and Inequality in European Union. Europolity-Continuity and Change in European Governance', 2020, 14(1), pp.5-31.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T. and Tsaneva-Atanasova, K., 2019. 'Artificial intelligence, bias and clinical safety', BMJ Quality & Safety, 28(3), pp.231-237.

Chaudhuri S, Long A, Zhang H, Monaghan C, Larkin JW, Kotanko P, et al. 'Artificial intelligence enabled applications in kidney disease', Semin Dial. 34:5–16, 2021.

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD, Krystal JH, Corlett PR. 'Cross-trial prediction of treatment outcome in depression: a machine learning approach', Lancet Psychiatry.;3(3):243-50, 2016.

Chi AC, Katabi N, Chen HS, Cheng YSL. Interobserver variation among pathologists in evaluating perineural invasion for oral squamous cell carcinoma. Head Neck Pathol. 10, 451–464, 2016.

Chinzei, K., Shimizu, A., Mori, K., Harada, K., Takeda, H., Hashizume, M., Ishizuka, M., Kato, N., Kawamori, R., Kyo, S. and Nagata, K., 'Regulatory science on AI-based medical devices and systems', Advanced Biomedical Engineering, 7, pp.118-123, 2018.

Chung Y, Addington J, Bearden CE, Cadenhead K, Cornblatt B, Mathalon DH, McGlashan T, Perkins D, Seidman LJ, Tsuang M, Walker E, Woods SW, McEwen S, van Erp TGM, Cannon TD; North American Prodrome Longitudinal Study (NAPLS) Consortium and the Pediatric Imaging, Neurocognition, and Genetics (PING) Study Consortium. Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. JAMA Psychiatr,75(9):960-968, 2018.

Clay H, Stern R. 'Making time in general practice', Primary Care Foundation, 1–83, 2015.

Cohen G. 'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?' Georgetown Law Journal. (108), 2020.

Collins GS, Moons KGM. 'Reporting of artificial intelligence prediction models', Lancet 393: 1577–79, 2019.

Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G. 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement', Circulation, 131(2), pp.211-219, 2015.

Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence. April 2021.

Cook C, Sheets C. 'Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy trials', J Man Manip Ther. 19(1):55-7, 2011.

Cook GJR, Goh V. What can artificial intelligence teach us about the molecular mechanisms underlying disease? Eur J Nucl Med Mol Imaging. 46:2715–2721, 2019.

Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, Rimm DL, Yang M, Romero E, Schalper KA, Velcheti V, Madabhushi A. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. Clin Cancer Res.;25(5):1526-1534, 2019.

Davenport, T.H., Barth, P. and Bean, R.,. How 'big data' Is different. MIT Sloan Management Review, 2012.

Dawoodbhoy FM, Delaney J, Cecula P, Yu J, Peacock I, Tan J, Cox B. AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. Heliyon. 7(5):e06993, 2021.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D. and van den Driessche, G., 'Clinically applicable deep learning for diagnosis and referral in retinal disease', Nature medicine, 24(9), pp.1342-1350, 2018.

De Vries L, Baselmans B, Bartels M. 'Smartphone-Based Ecological Momentary Assessment of Well-Being: A Systematic Review and Recommendations for Future Studies', Journal of Happiness Studies. 22:2361–2408, 2021.

Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB. On making the right choice: the deliberation-without-attention effect. Science. 2006; 311:1005e1007.

Directive, C., Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. Official Journal L, 210(07/08), pp.0029-0033, 1985.

Du, R., Lee, V. H., Yuan, H., Lam, K. O., Pang, H. H., Chen, Y., Lam, E. Y., Khong, P. L., Lee, A. W., Kwong, D. L., & Vardhanabhuti, V. 'Radiomics Model to Predict Early Progression of Non-metastatic Nasopharyngeal Carcinoma after Intensity Modulation Radiation Therapy: A Multicenter Study. Radiology', Artificial intelligence, 1(4), e180075, 2019.

Dusenbery, M. 'Everybody was telling me there was nothing wrong', The Health Gap, BBC News, 2018. www.bbc.com/future/article/20180523-how-gender-bias-affects-your-healthcare

Dwyer DB, Falkai P, Koutsouleris N. 'Machine learning approaches for clinical psychology and psychiatry', Annu Rev Clin Psychol, 14:91–118, 2018.

ECRIN, 'EFPIA, EATRIS, ELIXIR, BBMRI, ECRIN statement on the role of research infrastructures to boost patient-centred research and innovation in Europe', https://ecrin.org/news/efpia-eatris-elixir-bbmri-ecrin-statement-role-research-infrastructures-boost-patient-centred, 24 July 2019.

EGA Consortium (European Genome-Phenome Archive), https://ega-archive.org/datasets, 2021.

Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, et al. 'Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer' JAMA.;318(22):2199-2210., 2017.

Elements of AI. www.elementsofai.com, accessed November 2021.

Ellahham, S., Ellahham, N. and Simsekler, M.C.E., 'Application of artificial intelligence in the health care safety context: opportunities and challenges', American Journal of Medical Quality, 35(4), pp.341-348, 2020.

Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. 11(2):e1001603, 2014.

Emanuel EJ, Wachter RM. 'Artificial intelligence in health care: will the value match the hype?' JAMA. 321(23):2281-2282, 2019.

EuCanImage, https://eucanimage.eu, accessed November 2021.

European Commission, 'A Proposal for Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', April 2021.

European Commission. Digital Education Action Plan (2021-2027): Resetting Education for the Digital Age, 2020.

European Commission. Employment, Social Affairs & Inclusion Inequalities in access to healthcare. A study of national policies, 2018.

European Commission. The European Pillar of Social Rights in 20 principles. 2021.

European Genome-Phenome Archive, 'Browse datsasets', https://ega-archive.org/datasets, accessed November 2021.

European Health Data Space, https://ec.europa.eu/health/ehealth/dataspace_en, last access November, 2021.

Eurostat. Statistical expanded. Population structure and ageing, 2020.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J. 'A guide to deep learning in healthcare'. Nature medicine, 25(1), 24-29, 2019.

Evans AJ, Henry PC, Van der Kwast TH, Tkachuk DC, Watson K, Lockwood GA, Fleshner NE, Cheung C, Belanger EC, Amin MB, Boccon-Gibod L, Bostwick DG, Egevad L, Epstein JI, Grignon DJ, Jones EC, Montironi R, Moussa M, Sweet JM, Trpkov K, Wheeler TM, Srigley JR. 'Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens', Am J Surg Pathol. 32(10):1503-12, 2008.

Farina, R. and Sparano, A. 'Errors in sonography. In Errors in radiology' (pp. 79-85). Springer, Milano, 2012.

Felzmann, H., et al., 'Towards Transparency by Design for Artificial Intelligence,', Science and Engineering Ethics, 26:3333–3361, 2020.

Fernández García, J., Spatharou, A., Hieronimus, S., Beck, J.P., Jenkins, J. Transforming healthcare with AI: the impact on the workforce and organisations. Executive summary. EIT Health & McKinsey & Company, March 2020.

Ferryman, K. and Pitcan, M., Fairness in precision medicine. Data & Society, 2018.

Fihn SD, Saria S, Mendonça E, Hain S, Matheny M, Shah N, Liu H, Auerbach, A. 'Deploying AI in clinical settings. In artificial intelligence in health care: The hope, the hype, the promise, the peril', Editors: Matheny M, Israni ST, Ahmed M, Whicher D. Washington, DC: National Academy of Medicine, 2019.

Filice, R.W. and Ratwani, R.M. 'The case for user-centered artificial intelligence in radiology', Radiology: Artificial Intelligence, 2020, Vol. 2, No. 3.

Finlayson, S.G., Bowers, J.D. Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 'Adversarial attacks on medical machine learning', Science, 2019.

Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving. PubMed. Nat Biotechnol. 2018a.

Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, Osipov M, Kholodov M, Ismagilov R, Mohan S, Ostell J, Lu Z. 'Best Match: New relevance search for PubMed', PLoS Biol. 2018b;16(8):e2005343.

Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, Sarris J. 'The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials', World Psychiatry. 2017;16(3):287-298.

Fitzpatrick KK, Darcy A, Vierhile M. 'Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial', JMIR Ment Health.;4(2):e19, 2017.

Fleming N. 'How artificial intelligence is changing drug discovery', Nature: 557:S55–57, 2018.

Forster T, Kentikelenis A, Bambra, C, 'Health Inequalities in Europe: Setting the Stage for Progressive Policy Action', Foundation for European Progressive Studies. TASC: Think tank for action on social change. 2018.

Freeman, K, Dinnes, J, Chuchu, N, Takwoingi, Y, Bayliss, SE, Matin, RN, Jain, A, Walter, FM, Williams, HC and Deeks, JJ, 'Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies' BMJ, 2020, 368.

FUTURE-AI: Best practices for trustworthy AI in medical imaging, www.future-ai.eu, accessed November 2021.

Geis JR, Brady A, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R, Kotter E, Gichoya JW, Cook TS, Morgan MB, Tang A, Safdar NM, Kohli M. 'Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement', Insights Imaging, 2019 Oct 1;10(1):101. doi: 10.1186/s13244-019-0785-8. PMID: 31571015; PMCID: PMC6768929.

General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed December 2021.

Gerke, S., Minssen, T. and Cohen, G. 'Ethical and legal challenges of artificial intelligence-driven healthcare'. In Artificial intelligence in healthcare (pp. 295-336). Academic Press, 2020.

German Data Ethics Commission, Opinion of the Data Ethics Commission, July 2019, https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html.

Ghassemi, M. 'Exploring Healthy Models in ML for Health', AI for Healthcare Equity Conference, AI & Health at MIT, 2021. https://www.youtube.com/watch?v=5uZROGFYfcA

Gillespie, N., Lockey, S., & Curtis, C. 'Trust in Artificial Intelligence: A Five Country Study', The University of Queensland and KPMG Australia, 2021.

Gillies RJ, Kinahan PE, Hricak H. 'Radiomics: images are more than pictures, they are data,' Radiology.;278:563–577, 2016.

Giulietti M, Cecati M, Sabanovic B, Scirè A, Cimadamore A, Santoni M, et al. 'The role of artificial intelligence in the diagnosis and prognosis of renal cell tumors', Diagnostics, ;11(2):206, 2021.

Golbraikh A, Wang X, Zhu H, Tropsha A. 'Predictive QSAR modelling: methods and applications in drug discovery and chemical risk assessment', In Handbook of Computational Chemistry, ed. J Leszczynski, A Kaczmarek-Kedziera, T Puzyn, MG Papadopoulos, H Reis, MK, 2012.

Gómez-González E, Gómez E. 'Artificial Intelligence in medicine and healthcare: applications, availability and societal impact', EUR 30197 EN. Publications Office of the European Union, Luxembourg, 2020.

Goodfellow, I., Bengio, Y. and Courville, A., Deep learning. MIT Press, 2016.

Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, Jeste DV. 'Artificial intelligence for mental health and mental illnesses: An overview', Curr Psychiatry Rep;21:116, 2019.

Guo J, Li B. 'The application of medical artificial intelligence technology in rural areas of developing countries', Health Equity, ; 2: 174–81, 2018.

Gupta R, Kleinjans J and Caiment F. 'Identifying novel transcript biomarkers for hepatocellular carcinoma (HCC) using RNA-Seq datasets and machine learning', BMC Cancer.;21(962), 2021.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S. and Broderick, T., 'Transparency and reproducibility in artificial intelligence', Nature, 586(7829), pp.E14-E16, 2020.

Hamed S, Thapar-Björkert S, Bradby H, Ahlberg B. 'Racism in European Health Care: Structural Violence and Beyond', Sage Journals.;30(11), 2020.

Harned, Z., Lungren, M.P. and Rajpurkar, P. 'Machine vision, medical AI, and malpractice', Harv. JL & Tech. Dig, 2019.

Harvey, H.B. and Gowda, V., 'How the FDA regulates AI. Academic radiology', 27(1), pp.58-61, 2020.

Hashimoto DA, Rosman G, Witkowski ER, et al. 'Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy', Ann Surg.;270:414e421, 2019.

Hashimoto, D.A., Rosman, G., Rus, D. and Meireles, O.R., 'Artificial intelligence in surgery: promises and perils', Annals of surgery, 268(1), p.70, 2018.

Hermsen M, Bel T, Boer M Den, Steenbergen EJ, Kers J, Florquin S, et al. 'Deep learning-based histopathologic assessment of kidney tissue', J Am Soc Nephrol.;30(10):1968–79, 2019.

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J.P. and Shah, N.H. 'MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care', Journal of the American Medical Informatics Association, 27(12), pp.2011-2015, 2020.

Hill, N.R., Sandler, B., Mokgokong, R., Lister, S., Ward, T., Boyce, R., Farooqui, U. and Gordon, J., 'Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm', Journal of medical economics, 23(4), pp.386-393, 2020.

Hocking, L., Parks, S., Altenhofer, M. and Gunashekar, S.,. 'Reuse of health data by the European pharmaceutical industry', RAND Corporation, 2019.

Hoffman, K.M., Trawalter, S., Axt, J.R. and Oliver, M.N., 'Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites', Proceedings of the National Academy of Sciences, 113(16), pp.4296-4301, 2016.

Human-Centred Artificial Intelligence Programme, www.dtu.dk/english/Education/msc/Programmes/human-centered-artificial-intelligence, accessed November 2021.

Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. 'Prediction of sepsis patients using machine learning approach: a meta-analysis', Comput Methods Programs Biomed. 170:1-9, 2019.

Jamthikar AD, Gupta D, Saba L, Khanna NN, Viskovic K, Mavrogeni S, Laird JR, Sattar N, Johri AM, Pareek G, Miner M, Sfikakis PP, Protogerou A, Viswanathan V, Sharma A, Kitas GD, Nicolaides A, Kolluri R, Suri JS. 'Artificial intelligence framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound', Comput Biol Med.;126:104043, 2020.

Jiang S, Chin KS, Tsui KL. 'A universal deep learning approach for modeling the flow of patients under different severities', Comput Methods Programs Biomed.;154:191-203, 2018.

Jin, J.M., Bai, P., He, W., Wu, F., Liu, X.F., Han, D.M., Liu, S. and Yang, J.K.,. 'Gender differences in patients with COVID-19: focus on severity and mortality', Frontiers in public health, 8, p.152, 2020.

Kaddoum R, Fadlallah R, Hitti E, El-Jardali F, El Eid G. 'Causes of cancellations on the day of surgery at a Tertiary Teaching Hospital', BMC Health Serv. Res. 16, 2016.

Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F. 'Secure, privacy-preserving and federated machine learning in medical imaging', Nature Machine Intelligence, 2(6), pp.305-311, 2020.

Kamat AS, Parker A. 'Effect of perioperative inefficiency on neurosurgical theatre efficacy: a 15-year analysis', Br. J. Neurosurg. 29: 565–568, 2015.

Kaminski, M.E. and Malgieri, G. 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations. U of Colorado Law Legal Studies Research Paper', (19-28), 2019.

Kaushal, A., Altman, R. and Langlotz, C. 'Geographic distribution of US cohorts used to train deep learning algorithms', Jama, 324(12), pp.1212-1213, 2020.

Kiener, M. "You may be hacked' and other things doctors should tell you'. The Conversation. 3 November 2020. https://theconversation.com/you-may-be-hacked-and-other-things-doctors-should-tell-you-148946

Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y. and Park, S.H. 'Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers', Korean Journal of Radiology, 20(3), p.405, 2019.

Kirubarajan A, Taher A, Khan S, Masood S. 'Artificial intelligence in emergency medicine: A scoping review', J Am Coll Emerg Physicians Open.;1(6):1691-1702, 2019.

Koene, A., Clifton, C., Hatada, Y., Webb, H. and Richardson, R., A governance framework for algorithmic accountability and transparency, EPRS, European Parliament, 2019.

Kompa, B., Snoek, J. and Beam, A.L. 'Second opinion needed: communicating uncertainty in medical machine learning', NPJ Digital Medicine, 4(1), pp.1-6, 2021.

Koops, B.J.. 'The concept of function creep. Law, Innovation and Technology', 13(1), pp.29-56, 2021.

Krittanawong, C. 'The rise of artificial intelligence and the uncertain future for physicians', European Journal of Internal Medicine, 48, pp.e13-e14, 2018.

Kulkarni S, Seneviratne N, Baig MS, Khan AHA. 'Artificial Intelligence in Medicine: Where Are We Now?' Acad Radiol. Jan;27(1):62-70., 2020.

Kuo C-C, Chang C-M, Liu K-T, Lin W-K, Chiang H-Y, Chung C-W, et al. 'Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning', NPJ Digit Med.;2(29), 2019.

Lake IR, Colón-González FJ, Barker GC, Morbey RA, Smith GE, Elliot AJ. 'Machine learning to refine decision making within a syndromic surveillance service', BMC Public Health; 19: 559, 2019.

Larson, D.B., Harvey, H., Rubin, D.L., Irani, N., Justin, R.T. and Langlotz, C.P., 2021. 'Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: Summary and recommendations', Journal of the American College of Radiology, 18(3), pp.413-424, 2021.

Leavy, S. 'Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning', GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, May 2018.

Lee CS, Lee AY. 'How Artificial Intelligence Can Transform Randomized Controlled Trials', Transl Vis Sci Technol.;9(2):9, 2020.

Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. 'Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom', Biol Psychiatry Cogn Neurosci Neuroimaging.;6(9):856-864, 2021.

Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. 'Why digital medicine depends on interoperability', NPJ Digit Med.;2:79, 2019.

Lekadir, K. et al. 'FUTURE-AI: Best practices for trustworthy AI in medicine', www.future-ai.org, 2022.

Leone, D., Schiavone, F., Appio, F.P. and Chiao, B. 'How does artificial intelligence enable and enhance value co-creation in industrial markets? An exploratory case study in the healthcare ecosystem', Journal of Business Research, 129, pp.849-859, 2021.

Lewis, J.R. 'The system usability scale: past, present, and future', International Journal of Human–Computer Interaction, 34(7), pp.577-590, 2018.

Li, Y. and Vasconcelos, N. 'Repair: Removing representation bias by dataset resampling', In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9572-9581), 2019.

Lindenmeyer MT, Alakwaa F, Rose M, Kretzler M. 'Perspectives in systems nephrology,' Cell Tissue Res, 2021.

Lipton, Zachary C. 'The doctor just won't accept that!' arXiv preprint arXiv:1711.08037, 2017.

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis', Lancet Digit Health.1(6):e271-e297, 2019.

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J. and Denniston, A.K. 'Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension', BMJ, 370, 2020.

Loftus TJ, Filiberto AC, Li Y, Balch J, Cook AC, Tighe PJ, Efron PA, Upchurch GR Jr, Rashidi P, Li X, Bihorac A. 'Decision analysis and reinforcement learning in surgical decision-making', Surgery.168(2):253-266, 2020.

Loftus TJ, Upchurch GR Jr, Bihorac A. 'Use of Artificial Intelligence to Represent Emergent Systems and Augment Surgical Decision-Making', JAMA Surg. 154(9):791-792, 2019.

Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. 'Artificial Intelligence in Cardiology: Present and Future', Mayo Clin Proc; 95(5):1015–39, 2020.

Lorkowski J, Kolaszyńska O, Pokorski M. 'Artificial intelligence and precision medicine: a perspective', Adv Exp Med Biol. Jun 18. Doi. Epub ahead of print. PMID: 34138457, 2021.

Lundberg, S.M., Lee, S.I.. 'A unified approach to interpreting model predictions', in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4768–4777, 2017.

Lv, Z. and Piccialli, F. 'The security of medical data on Internet based on differential privacy technology', ACM Transactions on Internet Technology, 21(3), pp.1-18, 2021.

Maddox TM, Rumsfeld JS, Payne PRO. 'Questions for artificial intelligence in health care', JAMA. 321(1):31-32, 2019.

Madine, M.M., Battah, A.A., Yaqoob, I., Salah, K., Jayaraman, R., Al-Hammadi, Y., Pesic, S. and Ellahham, S. 'Blockchain for giving patients control over their medical records' IEEE Access, 8, pp.193102-193115, 2020.

Magrabi, F., Ammenwerth, E., McNair, J.B., De Keizer, N.F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P.J., Vehko, T., Wong, Z.S.Y. and Georgiou, A. 'Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implication', Yearbook of Medical Informatics, 28(1), p.128, 2019.

Maharana A, Nsoesie EO. 'Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity', JAMA Network Open. 1(4):e181535, 2018.

Maliha, G., Gerke, S., Cohen, I.G. and Parikh, R.B. 'Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation', The Milbank Quarterly, 2021.

Mamoshina P, Ojomoko L, Yanovich Y, Ostrovski A, Botezatu A, Prikhodko P, Izumchenko E, Aliper A, Romantsov K, Zhebrak A, Ogu IO, Zhavoronkov A. 'Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare', Oncotarget.;9:5665-5690, 2017.

Manne, R. and Kantheti, S.C. 'Application of artificial intelligence in healthcare: chances and challenges'. Current Journal of Applied Science and Technology, pp.78-89, 2021.

Marschang S, 'The European Health Data Space: is there room enough for all?' European Public Health Alliance', https://epha.org/the-european-health-data-space-is-there-room-enough-for-all/, 2021.

Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, Cook G. 'Introduction to Radiomics', J Nucl Med. 61:488-495, 2020.

McCoy, L.G., Nagaraj, S., Morgado, F., Harish, V., Das, S. and Celi, L.A. 'What do medical students actually need to know about artificial intelligence?' NPJ Digital Medicine, 3(1), pp.1-3, 2020.

McKeown, A., Mourby, M., Harrison, P., Walker, S., Sheehan, M. and Singh, I. 'Ethical issues in consent for the reuse of data in health data platforms', Science and Engineering Ethics, 27(1), pp.1-21, 2021.

McKinney, S. M. et al. 'International evaluation of an AI system for breast cancer screening', Nature 577, 89–94, 2020.

Medeiros J, Schwierz C. Efficiency estimates of health care systems in the EU. European Commission. Directorate-General for Economic and Financial Affairs. 2015.

Medeiros, J., Schwierz, C., 'Efficiency estimates of health care systems', Economic Papers, European Commission, 2015.

Menke NB, Caputo N, Fraser R, Haber J, Shields C, Menke MN. 'A retrospective analysis of the utility of an artificial neural network to predict ED volume', Am J Emerg Med.32:614-7, 2014.

Meskó B, Görög M. 'A short guide for medical professionals in the era of artificial intelligence', NPJ Digit Med. 3:126, 2020.

Michel JP, Ecarnot F. 'The shortage of skilled workers in Europe: its impact on geriatric medicine', Eur Geriatr Med. 11(3):345-347, 2020.

Miotto, R, Li L, Kidd BA, Dudley JIT. 'Deep patient: An unsupervised representation to predict the future of patients from the electronic health records', Scientific Reports. 6:26094, 2020.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD. 'Predicting cancer outcomes from histology and genomics using convolutional networks', Proc Natl Acad Sci U S A.; 115(13):E2970-E2979, 2018.

Mohr DC, Riper H, Schueller SM. 'A Solution-Focused Research Approach to Achieve an Implementable Revolution in Digital Mental Health', JAMA Psychiatry; 75(2):113-114, 2018.

Mooney SJ, Pejaver V. 'Big data in public health: terminology, machine learning, and privacy', Annu Rev Public Health;39:95-112, 2018.

Mora-Cantallops, M.; Sánchez-Alonso, S.; García-Barriocanal, E.; Sicilia, M.-A. 'Traceability for Trustworthy AI: A Review of Models and Tool', Big Data Cogn. Comput. 5, 20, 2021.

Morley, J. and Floridi, L. 'An ethically mindful approach to AI for health care', Lancet vol. 395, pp. 254-255, 2020.

Mulcahy, N. 'Recent Cyberattack Disrupted Cancer Care Throughout U.S' WebMD. 20 July 2021. https://www.webmd.com/cancer/news/20210720/recent-cyberattack-disrupted-cancer-care-us

Nagar A, Yew P, Fairley D, Hanrahan M, Cooke S, Thompson I, Elbaz W. 'Report of an outbreak of Clostridium difficile infection caused by ribotype 053 in a neurosurgery unit', J. Infect. Prev. 16: 126–130, 2015.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. 'Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies', BMJ. ;368:m689., 2020.

National Careers Service: The Skills Toolkit, https://nationalcareers.service.gov.uk/find-a-course/the-skills-toolkit, accessed November 2021.

Newman, L.H. 'These Hackers Made an App That Kills to Prove a Point', WIRED. 16 July 2019, https://www.wired.com/story/medtronic-insulin-pump-hack-app/.

NHS England. Clinical audit, https://www. england.nhs. uk/ clinaudit/, 2021.

NHS Improvement, Good Practice Guide: Focus on Improving Patient Flow, 2017. https://improvement.nhs.uk/documents/1426/Patient_Flow_Guidance_2017___13_July_2017.pdf

Niazi MKK, Parwani AV, Gurcan MN. 'Digital pathology and artificial intelligence', Lancet Oncol.; 20(5):e253-e261, 2019.

Noseworthy PA, Attia ZI, Brewer LPC, Hayes SN, Yao X, Kapa S, et al. 'Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis', Circ Arrhythmia Electrophysiol.;13(3), 2020.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S, 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

OECD/European Union, Health at a Glance: Europe 2020: State of health in the EU cycle. OECD Publishing, Paris, 2020.

OECD/European Union. Dementia prevalence. In Health at a Glance: Europe 2018: State of Health in the EU Cycle, OECD Publishing, Paris/European Union, Brussels, 2018.

Okanoue T, Shima T, Mitsumoto Y, Umemura A, Yamaguchi K, Itoh Y, Yoneda M, Nakajima A, Mizukoshi E, Kaneko S, Harada K. 'Artificial intelligence/neural network system for the screening of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis', Hepatol Res 51(5):554–569, 2021.

Ota, N., Tachibana, K., Kusakabe, T., Sanada, S. and Kondoh, M. 'A Concept for a Japanese Regulatory Framework for Emerging Medical Devices with Frequently Modified Behavior', Clinical and translational science, 13(5), pp.877-879, 2020.

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. 'A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images', Chaos, Solitons & Fractals, 140, 110190, 2020.

Paranjape, K., Schinkel, M., Panday, R.N., Car, J. and Nanayakkara, P. 'Introducing artificial intelligence training in medical education' JMIR Medical Education, 5(2), p.e16048, 2019.

Parikh RB, Teeple S, Navathe AS. 'Addressing bias in artificial intelligence in health care', JAMA; 322(24):2377-2378, 2019.

Park S, Park BS, Lee YJ, Kim IH, Park JH, Ko J, et al. 'Artificial intelligence with kidney disease: A scoping review with bibliometric analysis', PRISMA-ScR. Medicine (Baltimore);100(14), 2021.

Park, S.H. and Han, K. 'Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction', Radiology, 286(3), pp.800-809, 2018.

Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J. and Das, A.K. 'Evaluating artificial intelligence in medicine: phases of clinical research', Journal of American Medical Informatics Associations Open, 3(3), pp.326-331, 2020.

Peng J, Wang Y. 'Medical Image Segmentation with Limited Supervision: A Review of Deep Network Models', IEEE Access.; 99:, 2021.

Pérez MJ, Grande RG. 'Application of artificial intelligence in the diagnosis and treatment of hepatocellular carcinoma: A review', World J Gastroenterol. 26(37):5617–5628, 2021.

Pickering B. Trust, but Verify: Informed Consent, AI Technologies, and Public Health Emergencies, Future Internet 13(5):132, 2021.

Pinto, A., Pinto, F., Faggian, A., Rubini, G., Caranci, F., Macarini, L., Genovese, E.A. and Brunese, L. 'Sources of error in emergency ultrasonography', Critical Ultrasound Journal, 5(1), pp.1-, 2013.

Ploug, T, Holm S. 'Meta Consent –A Flexible Solution to the Problem of Secondary Use of Health Data', Bioethics, 30 (9), 2016.

Prokop M, van Everdingen W, van Rees Vellinga T, et al. 'CORADS— a categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation', Radiology, 2020:201473, [E-pub ahead of print, 2020 Apr 27].

Quaglio G, Brand H, Dario C. 'Fighting dementia in Europe: the time to act is now', Lancet Neurol. 15(5):452-4, 2016.

Quaglio GL, Boone R. What if we could fight drug addiction with digital technology?, EPRS, European Parliament, 2019.

Quaglio GL, Pirona A, Esposito G, Karapiperis T, Brand H, Dom G, Bertinato L, Montanari L, Kiefer F, Giuseppe Carrà G. 'Knowledge and utilization of technology-based interventions for substance use disorders: an exploratory study among health professionals in the European Union. Drugs: Education, Prevention and Policy; 26 (5): 437-446, 2018.

Quaglio GL. EU public health policy. 2020. European Parliamentary Research Services (EPRS). European Parliament, Brussels.

Quaglio GL, Millar S, Pazour M, Albrecht V, Vondrak T, Kwiek M, Schuch K. Exploring the performance gap in EU Framework Programmes between EU13 and EU15 Member States. 2020B. European Parliamentary Research Services (EPRS). European Parliament, Brussels.

Quer G, Arnaout R, Henne M, Arnaout R. 'Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review', J Am Coll Cardiol. 77(3):300–13, 2021.

Raghupathi, W. and Raghupathi, V. 'Big data analytics in healthcare: promise and potential. Health information science and systems', 2(1), pp.1-10, 2014.

Raji, I.D. 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', arXiv preprint, arXiv:2001.00973, 2020.

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H., 2018. 'Ensuring fairness in machine learning to advance health equity', Annals of Internal Medicine, 169(12), pp.866-872, 2018.

Ram S, Zhang W, Williams M, Pengetnze Y. 'Predicting asthma-related emergency department visits using big data', IEEE J Biomed Health Inform. 19:1216-23, 2015.

Rampton, V., Mittelman, M. and Goldhahn, J. 'Implications of artificial intelligence for medical education', The Lancet Digital Health, 2(3), pp.e111-e112, 2020.

Reardon, S. 'Rise of robot radiologists', Nature, 576(7787), pp.S54-S54, 2019.

Recht, M.P., Dewey, M., Dreyer, K., Langlotz, C., Niessen, W., Prainsack, B. and Smith, J.J. 'Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations', European Radiology, pp.1-9, 2020.

Redlich R, Almeida JJ, Grotegerd D, Opel N, Kugel H, Heindel W, et al. 'Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry—Pattern classification approach', JAMA Psychiatry; 71:1222–1230, 2014.

Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. 'Forecasting the onset and course of mental illness with Twitter data', Sci Rep. 7(1):13006, 2017.

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, 2015.

Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. 'Algorithmic impact assessments: A practical framework for public agency accountability', AI Now Institute, pp.1-22, 2018.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L. 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', AI & SOCIETY, pp.1-19, 2020.

Roski J, Chapman W, Heffner J, Trivedi R, Del Fiol G, Kukafka R, Bleicher Estiri OH, Klann J, Pierce J. 'How artificial intelligence is changing health and health care'. In Artificial Intelligence in Health Care: The hope, the hype, the promise, the peril. Editors: Matheny M, Israni ST, Ahmed M, Whicher D. Washington, DC: National Academy of Medicine, 2019.

Samulowitz A, Gremyr I, Eriksson E, Hensing G. ''Brave Men' and 'Emotional Women': A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain', Pain Res Manag. 2018;2018:6358624, 2018.

Sapci AH, Sapci HA. 'Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review', JMIR Aging ;2(2):e15429, 2019.

Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H.P., Janda, M., Condon, J.J., Oakden-Rayner, L., Palmer, L.J., Keel, S. and van Wijngaarden, P. 'A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology', Scientific Reports, 11(1), pp.1-10, 2021.

Schrider DR, Kern AD. 'Supervised machine learning for population genetics: a new paradigm', Trends Genet. 34:301–12, 2018.

Schwalbe N, Wahl B. 'Artificial intelligence and the future of global health', Lancet; 395(10236):1579-1586, 2020.

Schwartz WB. 'Medicine and the computer: the promise and problems of chang', N Engl J Med. 1970;283(23):1257-1264, 2020.

Scott, I., Carter, S. and Coiera, E. 'Clinician checklist for assessing suitability of machine learning applications in healthcare', BMJ Health & Care Informatics, 28(1), 2021.

Secretary-General of the OECD. Tackling wasteful spending on health, OECD Publishing, Paris, 2017.

Secretary-General of the OECD. Trustworthy AI in health. Background paper for the G20 AI Dialogue, Digital Economy Task Force, 2020.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y. and Ghassemi, M. 'CheXclusion: Fairness gaps in deep chest X-ray classifiers' BIOCOMPUTING 2021: Proceedings of the Pacific Symposium (pp. 232-243), 2021.

Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R. and Bakas, S. 'Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data', Scientific Reports, 10(1), pp.1-12, 2020.

Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. 'DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning' Sci Rep; 9:1879, 2019.

Shin EK, Mahajan R, Akbilgic O, Shaban-Nejad A. 'Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits', NPJ Digit Med; 1:50, 2018.

Shortliffe EH, Sepúlveda MJ. 'Clinical decision support in the era of artificial intelligence', JAMA.;320(21):2199-2200, 2018.

Shukla, 2016; pp. 2303–40. Dordrecht, Neth.: Springer.

Simpson S, Kay FU, Abbara S, et al. Radiological Society of North America Expert consensus statement on reporting chest CT findings related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA [E-pub ahead of print, 2020 Apr 28]. J Thorac Imaging 2020.

Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. 'Artificial intelligence-enhanced electrocardiography in cardiovascular disease management', Nat Rev Cardiol.;18:465–478, 2021.

Sit, C., Srinivasan, R., Amlani, A., Muthuswamy, K., Azam, A., Monzon, L. and Poon, D.S. 'Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey', Insights into Imaging, 11(1), p.14, 2020.

Smith, H. 'Clinical AI: opacity, accountability, responsibility and liability', AI & SOCIETY, pp.1-11, 2020.

Sornapudi S, Stanley RJ, Stoecker WV, Almubarak H, Long R, Antani S, Thoma G, Zuna R, Frazier SR. 'Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels', J Pathol Inform; 9:5, 2018.

Stanford University, Human-Centered Artificial Intelligence, https://hai.stanford.edu/, accessed November 2021.

Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. 'Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease.', PLoS One 13(8):e0202344, 2018.

Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, Thng F, Peng L, Stumpe MC.' Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer', Am J Surg Pathol. ;42(12):1636-1646, 2018.

Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. 'A Deep Learning Approach to Antibiotic Discovery', Cell. 180(4):688-702.e13, 2020.

Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A, Del Giudice C, Tesorio P, and Rusciano M. 'Precision and Personalized Medicine: How Genomic Approach Improves the Management of Cardiovascular and Neurodegenerative Disease', Genes. 11(7):747, 2020.

Stylianou N, Fackrell R, Vasilakis C. 'Are medical outliers associated with worse patient outcomes? A retrospective study within a regional NHS hospital using routine data', BMJ Open 7. e015676, 2017.

Subbaswamy, A. and Saria, S. 'From development to deployment: dataset shift, causality, and shift-stable models in health AI', Biostatistics, 21(2), pp.345-352, 2020.

Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, AP IJ, et al.' Advances and challenges in computational target prediction', J. Chem. Inf. Model. 59:1728–42, 2019.

Tanguay-Sela, M., Benrimoh, D., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., Fratila, R., Parikh, S., Karp, J., Heller, K. and Vahia, I. 'Evaluating the Usability and Impact of an Artificial Intelligence-Powered Clinical Decision Support System for Depression Treatment', Biological Psychiatry, 87(9), p.S171, 2020.

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment, ALTAI, European Commission, https://op.europa.eu/es/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1 2020.

The Assessment List for Trustworthy Artificial Intelligence, https://altai.insight-centre.org, accessed November 2021.

The World Bank, 'Maternal mortality ratio (modeled estimate, per 100,000 live births) – European Union', https://data.worldbank.org/indicator/SH.STA.MMRT?locations=EU, last accessed December 2021.

Tjoa, E. and Guan, C. 'A survey on explainable artificial intelligence (xai): Toward medical xai', IEEE Transactions on Neural Networks and Learning Systems, 2020.

Tlapa D, Zepeda-Lugo CA, Tortorella GL, Baez-Lopez YA, Limon-Romero J, Alvarado-Iniesta A, Rodriguez-Borbon MI. 'Effects of Lean Healthcare on Patient Flow: A Systematic Review', Value Health. 23(2):260-273, 2020.

Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. 'A clinically applicable approach to continuous prediction of future acute kidney injury', Nature. 572(7767):116–9, 2019.

Topol, EJ, 'High-performance medicine: the convergence of human and artificial intelligence' Nature Medicine, 25(1), 44–56, 2019.

TRIPOD, www.tripod-statement.org, accessed November 2021.

Tutt, A.. 'An FDA for algorithms', Admin. L. Rev., 69, p.83, 2017.

U.S. Food and Drug Administration (FDA). Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), 2019.

U.S. Food and Drug Administration (FDA), 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback'.

U.S. Food and Drug Administration (FDA). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, 2021

United Nations Educational, Scientific and Cultural Organization (UNESCO). Artificial Intelligence and Gender Equality: Key Findings of UNESCO's Global Dialogue, 2020.

United Nations News. 'More women and girls needed in the sciences to solve world's biggest challenges', February 2019.. https://news.un.org/en/story/2019/02/1032221

Viceconti, M, Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J. Musuamba Tshinanu, F. 'In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products', Methods 185; 120-127, 2021.

Vijayan V, Connolly J, Condell J, McKelvey N and Gardiner P. Review of Wearable Devices and Data Collection Considerations for Connected Health. Sensors. 2021; 21(16): 5589.

Vyas, D.A., et al. 'Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms', The New England Journal of Medicine (383), pp. 874-882., 2020.

Wager TD, Woo CW. 'Imaging biomarkers and biotypes for depression', Nat Med. 23(1):16-17, 2017.

Walsh CG, Ribeiro JD, Franklin JC. 'Predicting risk of suicide attempts over time through machine learning', Clin Psychol Sci; 5, 457–469, 2017.

Wanless D. 'Securing Good Health for the Whole Population', HM Treasury; 2004.

Westergaard, D., Moseley, P., Sørup, F.K.H., Baldi, P. and Brunak, S. 'Population-wide analysis of differences in disease progression patterns in men and women', Nature communications, 10(1), pp.1-14, 2019.

Whitby B. 'Automating medicine the ethical way', In: Pontier M (ed) Rysewyk Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering). Springer, Switzerland, 2015.

Wiggers, K. 'Google's breast cancer-predicting AI research is useless without transparency, critics say', VentureBeat, 14 October 2020. https://venturebeat.com/2020/10/14/googles-breast-cancer-predicting-ai-research-is-useless-without-transparency-critics-say/.

Williams, R. 'Lack of transparency in AI breast cancer screening study 'could lead to harmful clinical trials', scientists say', iNews UK, 14 October 2020.

Wolff, J., Pauling, J., Keck, A. and Baumbach, J. 'The economic impact of artificial intelligence in health care: systematic review', Journal of Medical Internet Research, 22(2), p.e16866, 2020.

Wood, A., Najarian, K. and Kahrobaei, D. 'Homomorphic encryption for machine learning in medicine and bioinformatics', ACM Computing Surveys (CSUR), 53(4), pp.1-35, 2020.

World Health Organization (WHO). Depression in Europe: facts and figures, 2021a. https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2012/10/depression-in-europe/depression-in-europe-facts-and-figures

World Health Organization (WHO). Ethics and governance of artificial intelligence for health: WHO guidance, 2021b.

World Health Organization (WHO). Global strategy on human resources for health: workforce 2030, Geneva, 2016. https://www.who.int/hrh/resources/pub_globstrathrh-2030/en/

Xu, W. 'Toward human-centered AI: a perspective from human-computer Interactions', 26(4), pp.42-46, 2019.

Yang, G., Ye, Q., & Xia, J. 'Unbox the Black box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Mini-Review, Two Showcases and Beyond' ArXiv, abs/2102.01998, 2021.

Yazdavar AH, Mahdavinejad MS, Bajaj G, Romine W, Sheth A, Monadjemi AH, Thirunarayan K, Meddar JM, Myers A, Pathak J, Hitzler P. 'Multimodal mental health analysis in social media', PLoS One; 15(4):e0226248, 2020.

Yu, K.H. and Kohane, I.S. 'Framing the challenges of artificial intelligence in medicine', BMJ Quality & Safety, 28(3), pp.238-241, 2019.

Zange L, Muehlberg F, Blaszczyk E, Schwenke S, Traber J, Funk S and Schulz-Menger J. 'Quantification in cardiovascular magnetic resonance: agreement of software from three different vendors on assessment of left ventricular function, 2D flow and parametric mapping', Journal of Cardiovascular Magnetic Resonance; 21:12, 2019.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K. 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study', PLoS Medicine, 15(11), p.e1002683, 2018.

Zhang BH, Lemoine B, Mitchell M. 'Mitigating unwanted biases with adversarial learning', In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335-340, 2018.

Zhang L, Tan J, Han D, Zhu H. 'From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today*; 22(11):1680–85, 2017.

Zhao L, Wang W, Sedykh A, Zhu H. 'Experimental errors in QSAR modeling sets: What we can do and what we cannot do', ACS Omega, 2:2805–12, 2017.

Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K. 'Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants', Chem. Res. Toxicol; 27:1643–51, 2014.

Zhu H. 'Big Data and Artificial Intelligence Modeling for Drug Discover', Annu Rev Pharmacol Toxicol. Jan 6;60:573-589, 2020.

QA-07-22-328-EN-N

In recent years, the use of artificial intelligence (AI) in medicine and healthcare has been praised for the great promise it offers, but has also been at the centre of heated controversy. This study offers an overview of how AI can benefit future healthcare, in particular increasing the efficiency of clinicians, improving medical diagnosis and treatment, and optimising the allocation of human and technical resources.

The report identifies and clarifies the main clinical, social and ethical risks posed by AI in healthcare, more specifically: potential errors and patient harm; risk of bias and increased health inequalities; lack of transparency and trust; and vulnerability to hacking and data privacy breaches.

The study proposes mitigation measures and policy options to minimise these risks and maximise the benefits of medical AI, including multi-stakeholder engagement through the AI production lifetime, increased transparency and traceability, in-depth clinical validation of AI tools, and AI training and education for both clinicians and citizens.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

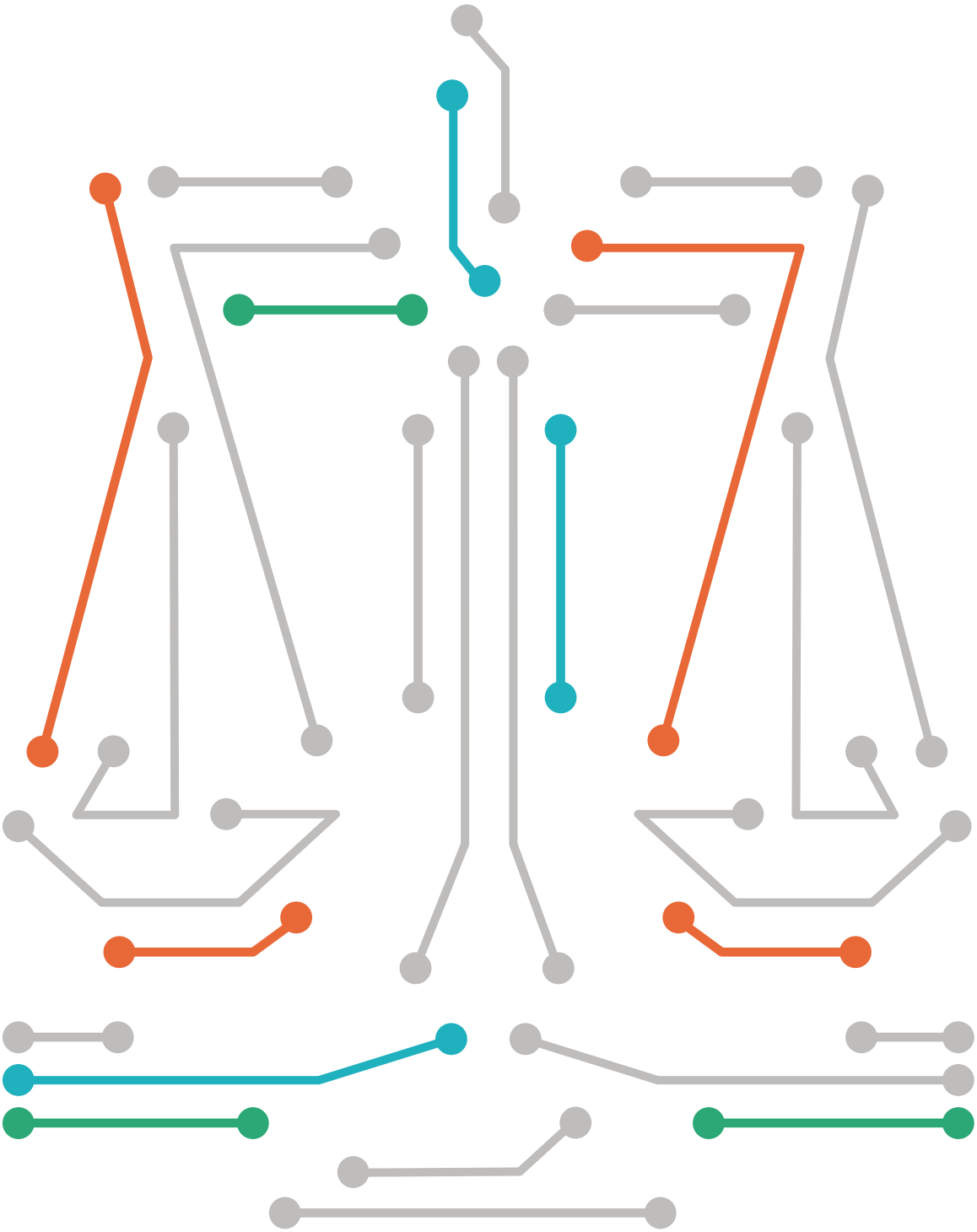


ISBN 978-92-846-9456-3 | doi: 10.2861/568473 | QA-07-22-328-EN-N



Regulatory considerations on artificial intelligence for health

World Health Organization



Regulatory considerations on artificial intelligence for health

World Health Organization

Regulatory considerations on artificial intelligence for health

ISBN 978-92-4-007887-1 (electronic version)
ISBN 978-92-4-007888-8 (print version)

© World Health Organization 2023

Some rights reserved. This work is available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo).

Under the terms of this licence, you may copy, redistribute and adapt the work for non-commercial purposes, provided the work is appropriately cited, as indicated below. In any use of this work, there should be no suggestion that WHO endorses any specific organization, products or services. The use of the WHO logo is not permitted. If you adapt the work, then you must license your work under the same or equivalent Creative Commons licence. If you create a translation of this work, you should add the following disclaimer along with the suggested citation: "This translation was not created by the World Health Organization (WHO). WHO is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition".

Any mediation relating to disputes arising under the licence shall be conducted in accordance with the mediation rules of the World Intellectual Property Organization (http://www.wipo.int/amc/en/mediation/rules/).

Suggested citation. Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see https://www.who.int/copyright.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

General disclaimers. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

Graphic design and layout by Phoenix Design Aid

CONTENTS

Foreword ..... **v**

Acknowledgements vii

Abbreviations and acronyms ix

Executive summary xi

1. Introduction 1

2. Purpose 2

3. Definitions, fundamental concepts and declarations of interest 3

4. Key artificial intelligence applications in health care and therapeutic development 4

5. Topic areas of regulatory considerations 6

5.1 Documentation and transparency **8**

5.1.1 Documentation across the total product lifecycle – ensuring a quality continuum 9

5.1.2 Pre-specification and documenting the medical purpose, clinical context and development 10

5.1.3 Deployment and post-deployment 10

5.1.4 Risk-based approach and proportionality 10

5.2 Risk management and artificial intelligence systems development lifecycle approach **12**

5.2.1 AI systems during the development and deployment process 12

5.2.2 AI systems development lifecycle 13

5.2.3 Holistic risk management 14

5.3 Intended use and analytical and clinical validation **20**

5.3.1 Use case description, analytical and clinical validation 20

5.3.2 Intended use 21

5.3.3 Analytical validation (also referred to as technical validation) 22

5.3.4 Clinical validation 24

5.3.5 Post-market monitoring 24

5.3.6 Changes to the AI tool 25

5.3.7 Low- and middle-income countries 25

5.4 Data quality **27**

5.4.1 Data in current health ecosystems 27

5.4.2 Good quality data in health AI systems 27

5.4.3 Key quality data challenges and considerations for health AI systems 27

5.5 Privacy and data protection 33
5.5.1 Current landscape 33
5.5.2 Documentation and transparency 35
5.5.3 AI regulatory sandboxes 37
5.6 Engagement and collaboration 39
5.6.1 Discussion on strategies of profiled regulatory bodies 43
5.6.2 Two successful instances of engagement 44
5.6.3 Recommended approaches for countries without past experience 45
5.6.4 Narrative on using engagement tools based on practical experience 45
5.6.5 Narrative positioning the regulator as a partner in the development process 46
6. Recommendations for the way forward 48
7. Conclusion 50
References 51
Annex. Definitions, fundamental concepts and declarations of interest 58

FOREWORD

The mission of the World Health Organization (WHO) to promote health, keep the world safe and serve the vulnerable is articulated in its global strategy on digital health 2020–2025. At the heart of this strategy, WHO aims to improve health for everyone, everywhere by accelerating the development and adoption of appropriate, accessible, affordable, scalable and sustainable person-centric digital health solutions to prevent, detect and respond to epidemics and pandemics, developing infrastructure and applications. WHO – along with many international and regional organizations and national authorities – recognizes the potential of Artificial Intelligence (AI) in accelerating the digital transformation of health care. AI has an evident potential to strengthen health service delivery to underserved populations, enhance public health surveillance, advance health research and the development of medicines, support health systems management and enable clinical professionals to improve patient care and perform complex medical diagnoses. However, existing and emerging AI technologies, including large language models, are being rapidly deployed without a full understanding of how such AI systems may perform – potentially either benefitting or harming end-users, including health-care professionals and patients.

Consequently, to facilitate the safe and appropriate use of AI technologies for the development of AI systems in health care, the WHO and the International Telecommunication Union (ITU) have established a Focus Group on AI for Health (FG-AI4H). To support its work, FG-AI4H has created several working groups, including a Working Group on Regulatory Considerations (WG-RC) on AI for Health. The WG-RC consists of members representing multiple stakeholders – including regulatory authorities, policy-makers, academia and industry – who have explored regulatory and health technology assessment concepts and emerging “good practices” for the development and use of AI in health care and therapeutic development. The work of the WG-RC represents a multidisciplinary, international effort to increase dialogue and examine key considerations on the use of AI in health care.

This document provides an overview of regulatory considerations on AI for health that covers key general topic areas, namely: documentation and transparency, risk management and AI systems development lifecycle approach, intended use and analytical and clinical validation, AI related data quality, privacy and protection, and engagement and collaboration. In addition, the publication recommends that stakeholders take into account 18 regulatory considerations as they continue to develop frameworks and best practices for the use of AI in health care. This document is intended to be a listing of key regulatory considerations and as a resource that can be considered by all relevant stakeholders in medical devices ecosystems, including developers who are exploring and developing AI systems, regulators who might be in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI-embedded medical devices, health practitioners who deploy and use such medical devices and AI systems, and others working in these areas. The document recommends that stakeholders examine the key considerations and other potential ones as they continue to develop frameworks and best practices for the use of AI in health care in relationship to the key topic areas.

I wish to thank all subgroup technical experts, external expert group members, external reviewers, stakeholders, and partners in the United Nations family and beyond who made essential contributions to the development of this document. I hope that this report will help to ensure that the development and use of AI for health and

will be guided by appropriate regulatory considerations so that all populations can safely and effectively benefit from the great promise these technologies hold for the future.





Dr. Jeremy Farrar, Chief Scientist, World Health Organization

ACKNOWLEDGEMENTS

Development of this document was led by Sameer Pujari (Department of Digital Health and Innovation) and Shada AlSalamah (Department of Digital Health and Innovation) under the overall guidance of Alain Labrique (Director, Department of Digital Health and Innovation), Jeremy Farrar (Chief Scientist, Science Division), Soumya Swaminathan (Former Chief Scientist, Science Division), Bernardo Mariano (Former Director, Department of Digital Health and Innovation), Adriana Velazquez Berumen (Access to Medicines and Health Products), and Anita Sands (Regulation and Prequalification Department).

Technical coordination of the topic areas was provided by the following subgroup leads (in alphabetical order): Shada Alsalamah (Department of Digital Health and Innovation) led the subgroup on the Data Quality and Risk Management and AI Systems Development Lifecycle Approaches; M Khair ElZarrad, (U.S. Food and Drug Administration, United States of America) led the Documentation and Transparency topic area; Monique Kuglitsch (Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Germany) and Dean Ho (National University of Singapore, Singapore) jointly led the Engagement and Collaboration topic area; Naomi Lee (Former The Lancet, United Kingdom of Great Britain and Northern Ireland) led the Intended Use and Analytical and Clinical Validation topic area; and Rose Purcell (Former U.S. Food and Drug Administration, United States of America) led the Privacy and Data Protection topic area.

WHO are grateful to the following persons who contributed to development and review of this publication (in alphabetical order): Najeeb Al-Shorbaji (eHealth Development Association, Jordan), Batoul Albaz (Health Sector, King Abdulaziz City for Science and Technology, Saudi Arabia), Paolo Alcini (European Medicines Agency, Netherlands (Kingdom of the)), Ali AlDalaan (Saudi Food and Drug Administration, Saudi Arabia), Fazilah Shaik Allaudin (Ministry of Health, Malaysia), Safaa Dirar Almajthoub (Digital Health Center of Excellence, Ministry of Health, Saudi Arabia), Asma Ibrahim Al Mannaei (Drugs & Medical Products Division, Department of Health- Abu Dhabi, United Arab Emirates), Abdulgader Almoeen (National Center for AI, Saudi Data & AI Authority, Saudi Arabia), Sultan Alzahrani (Digital Health Institute, Health Sector, King Abdulaziz City for Science and Technology, Saudi Arabia), Lin Anle (Health Sciences Authority, Singapore), Pat Baird (FG-AI4H, United States of America), Michael Berensmann (Federal Institute for Drugs and Medical Devices, Germany), Simão de Campos Neto (ITU), Mattias Karlsson Dinnetz (IP for Innovators Department, Technology Transfer Section, World Intellectual Property Organization), Tala H. Fakhouri (U.S. Food and Drug Administration, United States of America), Hélio Bomfim de Macêdo Filho (Brazilian Health Regulatory Agency- Anvisa, Brazil), Luca Foschini (FG-AI4H, United States of America), Mohammed VI Hassan Ghazal (University of Health Sciences, Morocco), Liang Hong (China's Center for Medical Device Evaluation, China), Indra Joshi (NHSX, National Health Service, United Kingdom of Great Britain and Northern Ireland), Kassandra Karpathakis (Harvard TH Chan School of Public Health, United States of America; NHSX, National Health Service, United Kingdom of Great Britain and Northern Ireland), Tim Kelsey (Healthcare Information and Management Systems Society, United Kingdom of Great Britain and Northern Ireland), Andrea Keyter (FG-AI4H, South Africa), Vladimir Kutichev (Russian Federal Service for Surveillance in Healthcare, Russian Federation), Marc Lamoureux (Health Canada, Canada), Tze-Yun Leong (National University of Singapore and AI Singapore, Singapore), Xiaoxuan Liu (University of Birmingham, United Kingdom of Great Britain and Northern Ireland), Junaid Nabi (Harvard Business School, United States of America), Mariam Nouh (Future Economies Sector, King Abdulaziz City for Science and Technology, Saudi Arabia), Luis Oala (Fraunhofer Heinrich Hertz Institute, Germany), Mats Ohlson (FG-AI4H, Sweden), Adrian Pacheco-Lopez (National Center for

Health Technology Excellence, Ministry of Health, Mexico), Ugo Pagallo (University of Turin, Italy), Maria Beatrice Panico (Medicines and Healthcare products Regulatory Agency, United Kingdom of Great Britain and Northern Ireland), Andres Pichon-Riviere (Institute for Clinical Effectiveness and Health Policy, Argentina), Julie Polisena (Health Canada, Canada), Pierre Quartarolo (Danish Medicines Agency, Denmark), Chandrashekar Ranga (The Central Drugs Standard Control Organisation, India), Mansooreh Saniei (King's College London, United Kingdom of Great Britain and Northern Ireland), Raymond Francis R. Sarmiento (University of the Philippines Manila, Philippines), Kanako Sasaki (Ministry of Health, Labour and Welfare, Japan), Brian Scarpelli (FG-AI4H, United States of America), Rama Sethuraman (Health Sciences Authority, Singapore), Robert Ssekitoleko (Makerere University, Uganda), Bev Townsend (University of York, United Kingdom of Great Britain and Northern Ireland), Tayab Waseem (FG-AI4H, United States of America), Thomas Wiegand (Fraunhofer Heinrich Hertz Institute, Germany), and Georg Zimmermann (Paracelsus Medical University, Austria).

Additional contributions were also received from the following WHO staff (in alphabetical order): Samvel Azatyan, Judith Van Andel (former), Housseynou Ba, Mengjuan Duan, Marcelo D'Agostino, Jose Eduardo Díaz Mendoza, Clayton Hamilton, Josee Hansen, Wouter 'T Hoen, Agnes Sitta Kijo, Rohit Malpani, Ahmed Mandil, Mohamed Nour, David Novillo Ortiz, Andreas Reis, Denise Schalet, Mariam Shokralla (former) and Yu Zhao.

ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
CDSS	Clinical Decision Support System
CONSORT-AI	Consolidated Standards of Reporting Trials for AI
CQC	Care Quality Commission
CRM-N	Clinical Research Materials Notification
DAISAM	Data and artificial intelligence assessment methods
DHSC	Department of Health and Social Care
EC	European Commission
EU	European Union
FG-AI4H	Focus Group on Artificial Intelligence for Health
GDPR	General Data Protection Regulation
GHWP	Global Harmonization Working Party
HIPAA	Health Insurance Portability and Accountability Act
HSA	Health Sciences Authority
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICMRA	International Coalition of Medicines Regulatory Authorities
iDAIR	The International Digital Health & AI Research Collaborative
IMDRF	International Medical Device Regulators Forum
IoT	Internet of Things
IP	Intellectual property
ISO	International Organization for Standardization
ITU	International Telecommunication Union
MAS	Multi-agent systems
MHRA	Medicines and Healthcare Products Regulatory Agency
ML	Machine learning
NHS	National Health Service

NICE	National Institute for Health and Care Excellence
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
PACMP	Post-Approval Change Management Protocol
PMDA	Japanese Pharmaceuticals and Medical Devices Agency
QMS	Quality management system
SAHPRA	South African Health Products Regulatory Authority
SaMD	Software as a Medical Device
SANAS	South African National Accreditation System
SAR	Special access route
SPIRIT-AI	Standard Protocol Items: Recommendations for Interventional Trials for AI
TGA	Therapeutic Goods Administration
TPLC	Total Product Lifecycle
US FDA	U.S. Food and Drug Administration
WG-RC	Working Group on Regulatory Considerations on Artificial Intelligence for Health
WHO	World Health Organization
WIPO	World Intellectual Property Organization

EXECUTIVE SUMMARY

The mission of the World Health Organization (WHO) is to promote health, keep the world safe and serve the vulnerable is articulated in its global strategy on digital health 2020–2025 *(1)*. At the heart of this strategy, WHO aims to improve health for everyone, everywhere by accelerating the development and adoption of appropriate, accessible, affordable, scalable and sustainable person-centric digital health solutions in order to prevent, detect and respond to epidemics and pandemics, developing infrastructure and applications. Many international organizations and global players are contributing to this area along with WHO, including The International Medical Device Regulators Forum (IMDRF), Global Harmonization Working Party (GHWP), the US Food and Drug Administration (U.S. FDA), Health Canada, the International Coalition of Medicines Regulatory Authorities (ICMRA), the International Organization for Standardization (ISO), the Organisation for Economic Co-operation and Development (OECD), the United Kingdom of Great Britain and Northern Ireland's Medicines and Healthcare Products Regulatory Agency (MHRA), the South African Health Products Regulatory Authority (SAHPRA), the European Commission (EC), Singapore's Health Sciences Authority (HSA), the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), Japan's Pharmaceuticals and Medical Devices Agency (PMDA), Swissmedic and Australia's Therapeutic Goods Administration (TGA). These international and regional organizations and national authorities collectively recognize the potential of Artificial Intelligence (AI) in enhancing health outcomes by improving clinical trials, medical diagnosis and treatment, self-management of care and personalized care, as well as by creating more evidence-based knowledge, skills and competencies for professionals to support health care. Furthermore, with the increasing availability of health-care data and the rapid progress of analytics techniques, AI has the potential to transform the health sector to meet a variety of stakeholders' needs in health care and therapeutic development.

In order to facilitate the safe and appropriate use of AI technologies for the development of AI systems in health care, the WHO and the International Telecommunication Union (ITU) have established a Focus Group on AI for Health (FG-AI4H). To support its work, FG-AI4H created several working groups, including a Working Group on Regulatory Considerations (WG-RC) on AI for Health. The WG-RC consists of members representing multiple stakeholders – including regulatory authorities, policy-makers, academia and industry – who explored regulatory and health technology assessment concepts and emerging "good practices" for the development and use of AI in health care and therapeutic development. The work of the WG-RC represents a multidisciplinary, international effort to increase dialogue and examine key concepts for the use of AI in health care.

This publication, which is based on the work of the WG-RC, aims to deliver an overview of regulatory considerations on AI for health that covers the following six general topic areas: documentation and transparency, the total product lifecycle approach and risk management, intended use and analytical and clinical validation, data quality, privacy and data protection, and engagement and collaboration. This overview is not intended as guidance or as a regulatory framework or policy. Rather, it is a discussion of key regulatory considerations and a resource that can be considered by all relevant stakeholders – including developers who are exploring and developing AI systems, regulators and policy-makers who in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI-enabled medical devices, and health practitioners who deploy and use such medical devices and AI systems. Consequently, the WG-RC recommends that stakeholders take into account the following considerations as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development:

1. **Documentation and transparency:** Pre-specifying and documenting the intended medical purpose and development process – such as the selection and use of datasets, reference standards, parameters, metrics, deviations from original plans and updates during the phases of development – should be considered in a manner that allows for the tracing of the development steps as appropriate. A risk-based approach should be considered also for the level of documentation and record-keeping utilized for the development and validation of AI systems.

2. **Risk management and AI systems development lifecycle approaches:** A total product lifecycle approach should be considered throughout all phases in the life of an AI system, namely: pre-market development management, post-market surveillance and change management. In addition, it is essential to consider a risk management approach that addresses risks associated with AI systems, such as cybersecurity threats and vulnerabilities, underfitting, algorithmic bias etc.

3. **Intended use, and analytical and clinical validation:** Initially, providing transparent documentation of the intended use of the AI system should be considered. Details of the training dataset composition underpinning an AI system – including size, setting and population, input and output data and demographic composition – should be transparently documented and provided to users. In addition, it is key to consider demonstrating performance beyond the training and testing data through external analytical validation in an independent dataset. This external validation dataset should be representative of the population and setting in which it is intended to deploy the AI system and should be independent of the dataset used for developing the AI model during training and testing. Transparent documentation of the external dataset and performance metrics should be provided. Furthermore, it is important to consider a graded set of requirements for clinical validation based on risk. Randomized clinical trials are the gold standard for evaluation of comparative clinical performance and could be appropriate for the highest-risk tools or where the highest standard of evidence is required. In other situations, prospective validation can be considered in a real-world deployment and implementation trial which includes a relevant comparator that uses accepted groups. Finally, a period of more intense post-deployment monitoring should be considered through post-market surveillance and market surveillance for AI systems.

4. **Data quality:** Developers should consider whether available data are of sufficient quality to support the development of the AI system to achieve the intended purpose. Furthermore, developers should consider deploying rigorous pre-release evaluations for AI systems to ensure that they will not amplify any of the issues discussed in Section 5.4 of this document, such as biases and errors. Careful design or prompt troubleshooting can help identify data quality issues early and can prevent or mitigate possible resulting harm. Stakeholders should also consider mitigating data quality issues and the associated risks that arise in health-care data, as well as continue to work to create data ecosystems to facilitate the sharing of good-quality data sources.

5. **Privacy and data protection:** Privacy and data protection should be considered during the design and deployment of AI systems. Early in the development process, developers should consider gaining a good understanding of applicable data protection regulations and privacy laws and should ensure that the development process meets or exceeds such legal requirements. It is also important to consider implementing a compliance programme that addresses risks and ensures that the privacy and cybersecurity practices take into account potential harm as well as the enforcement environment.

6. **Engagement and collaboration:** During development of the AI innovation and deployment roadmap it is important to consider the development of accessible and informative platforms that facilitate

engagement and collaboration among key stakeholders, where applicable and appropriate. It is fundamental to consider streamlining the oversight process for AI regulation through such engagement and collaboration in order to accelerate practice-changing advances in AI.

Finally, the WG-RC has provided a forum for regulators and subject matter experts to discuss regulatory considerations for the use of AI technologies and development of AI systems for health and medical purposes. The WG-RC recognizes that the AI landscape is evolving rapidly and that the considerations in this deliverable may require expansion as technology and its uses develop. The working group recommends that stakeholders, including regulators and developers, continue to engage and that the community at large works towards shared understanding and mutual learning. In addition, established national and international groups, such as the International Medical Device Regulators Forum (IMDRF) and the International Coalition of Medicines Regulatory Authorities (ICMRA) should continue to work on topics of AI for potential regulatory convergence and harmonization.

1. INTRODUCTION

The mission of the World Health Organization (WHO) to promote health, keep the world safe and serve the vulnerable is articulated in its Global strategy on digital health 2020–2025 *(1)*. At the heart of this strategy, WHO aims to improve health for everyone, everywhere by accelerating the development and adoption of appropriate, accessible, affordable, scalable and sustainable person-centric digital health solutions to prevent, detect and respond to epidemics and pandemics. This should enable countries to use health data to promote health and well-being in order to achieve the United Nation's health-related Sustainable Development Goals (SDGs) *(2)* and the triple billion targets of WHO's Thirteenth General Programme of Work, 2019–2023 *(3)*.

In addition to WHO's efforts, there is a wave of interest by many other international and regional organizations. Key players include the International Medical Device Regulators Forum (IMDRF) *(4)*, the Global Harmonization Working Party (GHWP), the International Coalition of Medicines Regulatory Authorities (ICMRA) *(5)*, the International Organization for Standardization (ISO) *(6)*, the Organisation for Economic Co-operation and Development (OECD) *(7)* and the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH). Moreover, there are national efforts sharing the same goal.[1]

The digital transformation of health care and therapeutic development, which includes exploring the uses of Artificial Intelligence (AI), has the potential to enhance health outcomes by improving medical diagnosis, digital therapeutics, clinical trials, self-care and evidence-based knowledge. For the purpose of this document AI is defined as "a branch of computer science, statistics, and engineering that uses algorithms or models to perform tasks and exhibit behaviors such as learning, making decisions and making predictions. The subset of AI known as Machine Learning (ML) allows computer algorithms to learn through data, without being explicitly programmed, to perform a task" *(8)*. With the increasing availability of health-care data and the rapid progress in analytics techniques, AI has the potential to transform the health sector, which is one of the most important sectors for societies and economies worldwide.

[1] A non-exclusive list of national efforts: US Food and Drug Administration (US FDA), Health Canada, the Medicines and Healthcare Products Regulatory Agency (MHRA) of the United Kingdom of Great Britain and Northern Ireland, the South African Health Products Regulatory Authority (SAHPRA), the European Commission (EC), the Singapore Health Sciences Authority (HSA), Japan's Pharmaceuticals and Medical Devices Agency (PMDA), Swissmedic and Australia's Therapeutic Goods Administration (TGA).

2. PURPOSE

The International Telecommunication Union (ITU) is the United Nation's specialized agency for information and communications technology while WHO is the United Nation's specialized agency for health. These organizations partnered to establish an open group of experts to develop a generalizable benchmarking[2] framework for health solutions based on AI – the ITU/WHO Focus Group on AI for Health (FG-AI4H). In order to facilitate the safe and appropriate use of AI technologies[3] for the development of AI systems[4] in health care and support its work, the FG-AI4H created a Working Group on Regulatory Considerations (WG-RC) on AI for Health. The WG-RC consists of multiple stakeholders – including representatives from regulatory authorities, policy-makers, academia and industry – who explored regulatory and health technology assessment concepts and emerging "good practices" for the development and use of AI in health care and therapeutic development.

This publication is a general, high-level and nonexclusive overview of key regulatory considerations in topic areas developed by the WG-RC to support the overarching FG-AI4H framework. Recognizing that a single publication cannot address the specifics of the various AI systems that can be used for therapeutic development or health-care applications in general, the WG-RC's overview will highlight some of the key regulatory principles and concepts – such as risk–benefit assessments and considerations for the evaluation and monitoring of the performance of AI systems developed using AI technologies. Throughout the process of developing this publication, the WG-RC took into consideration different stakeholder perspectives, as well as different global and regional settings. The WG-RC's overview is not intended as guidance, as a regulatory framework or policy. Rather, it is meant as a listing of key regulatory considerations and a resource for all relevant stakeholders – including developers who are exploring and using AI technologies and developing AI systems, regulators who might be in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI systems that are embedded in medical devices, and health practitioners who deploy and use such medical devices and AI systems.

[2] This framework should not be confused with WHO's global benchmarking tool for the evaluation of national regulatory systems (https://www.who.int/tools/global-benchmarking-tools, accessed 25 July 2023).

[3] In the context of this publication, the term "AI technology" is used to refer to any AI technology (e.g. machine learning, deep learning, natural language processing, computer vision etc.) that is used to develop an AI system.

[4] An AI system is an AI-based system that is able to perform tasks such as visual perception, speech recognition, decision-making and translation between languages by using machine learning (ML) (including deep learning) or non-ML expert systems (based on rules such as decision trees). For example, an ML-enabled medical device uses ML, in part or in whole, to achieve its intended medical purpose and can therefore be considered an AI-based system.

3. DEFINITIONS, FUNDAMENTAL CONCEPTS AND DECLARATIONS OF INTEREST

For the purpose of this document, some key terms and concepts are defined and/ or explained in the Annex, as is the approach used to assess and manage external participants' declarations of interest.

4. KEY ARTIFICIAL INTELLIGENCE APPLICATIONS IN HEALTH CARE AND THERAPEUTIC DEVELOPMENT

AI is increasingly being explored to advance health care on multiple fronts. The blending of technology and medicine in research and development is facilitating a wealth of innovation that continues to improve *(9)*. Many health-related AI systems already exist or are being developed to meet a variety of stakeholders' needs in health care and therapeutic development. These solutions have wide-ranging uses across the spectrum of health-care delivery and therapeutic development. For instance, AI systems are being used in health care to support patients throughout the diagnosis and treatment of a disease, using solutions that support adherence to therapeutics and enhance communication capabilities with care providers.

Health care is becoming more patient-centric with personalized approaches to decision-making. This allows data to be used to improve patient and population wellness, patient education and engagement, prevention and prediction of diseases and care risks, medication adherence, disease management, disease reversal/remission, and individualization and personalization of treatment and care. Toward these ends, AI is increasingly being incorporated and utilized in the clinical setting. For instance, AI-enabled medical devices are being utilized to support clinical decision-making, and AI systems can facilitate clinical assessment of patients and care triaging. AI systems are also being used in the development and evaluation of medical products, including during drug discovery to identify potential therapeutic candidates and in clinical research for patient enrichment. Figure 1 illustrates areas of AI research and development across the spectrum of health-care delivery and therapeutic development. The figure does not show an exhaustive listing of all AI applications but instead provides examples that are meant to show the broad range of current and potential uses of AI systems.

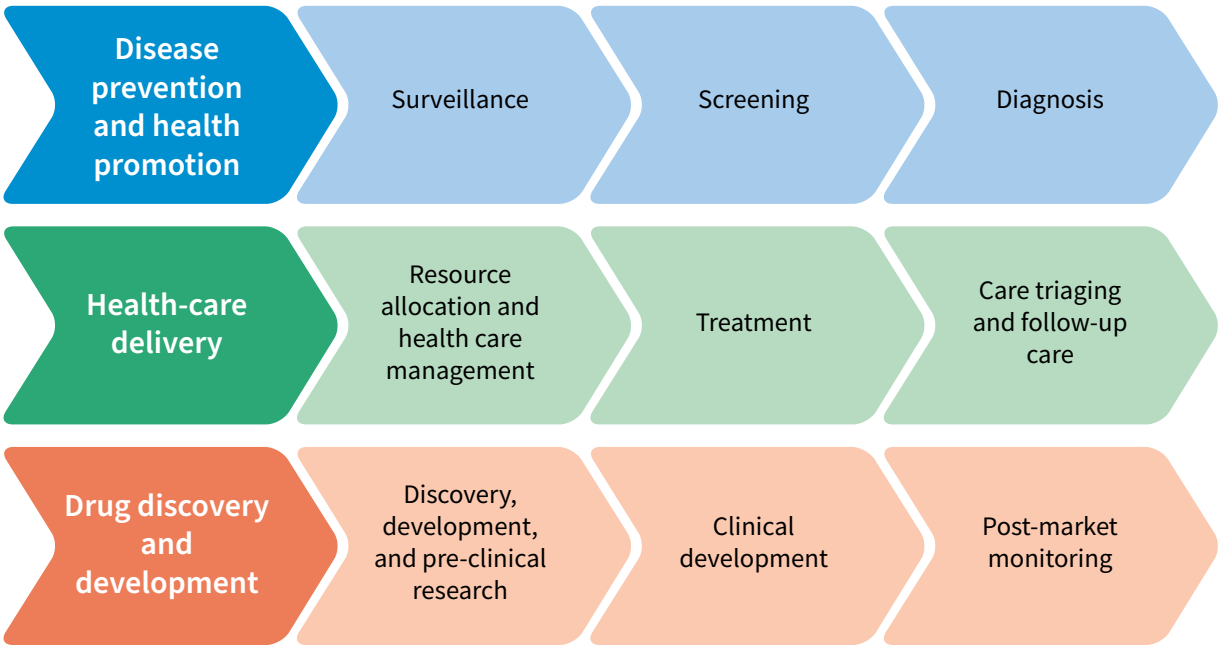


FIGURE 1. A general spectrum of AI research and development in health-care delivery and therapeutic development

The spectrum in Figure 1 assists in determining what regulatory considerations may be applicable and how they can be implemented. This document describes a selection of key regulatory considerations and discusses topic areas that are relevant to many stakeholders in the current AI for health ecosystem.

5. TOPIC AREAS OF REGULATORY CONSIDERATIONS

AI systems may be utilized across all aspects of health care and therapeutic development. Regardless of the category of the AI system application, regulators are keen to ensure not only that the AI systems are safe and effective for intended use but also that such promising tools reach those who need them as fast as possible. Dialogue between all stakeholders participating in the AI for health ecosystem – especially developers, manufacturers, regulators, users and patients – is highly advised as the AI community matures. Consequently, this publication aims to establish a common understanding of the use of AI systems in health that can be relevant to stakeholders.

The topic areas' subgroup leads performed a systematic literature review in 2020 of scientific publications in PubMed and other databases which included current guidelines and good practices in health care and therapeutic development. These sources were considered to define the list of topic areas of regulatory considerations for the use of AI in health care and therapeutic development. At its first meeting, the WG-RC discussed the proposed topic areas and sought consensus to focus its deliverable on the six key areas listed in Table 1 while also discussing the remaining sections of this publication. The working group was divided into six subgroups composed of subject matter experts who drafted a section on each topic area.

The WG-RC stressed that this list is not a fully inclusive list of key considerations. The working group expects that the list will serve as a starting point for future deliberations and subsequent updates. For example, global systems for protecting intellectual property (IP) may be an important area to discuss as part of cross-jurisdiction regulations for some stakeholders (mainly AI system developers and manufacturers), and also in relation to, for instance, the protection of AI-related inventions by way of laws on patents and trade secrets. Although not addressed in this report, the World Intellectual Property Organization (WIPO) has already begun a dialogue on AI and IP *(10).* Thus, WHO will engage in this work together with WIPO and other relevant stakeholders.

TABLE 1. Six key topic areas of regulatory considerations

Topic Area No.	Topic Area Name
Topic Area 1	Documentation and transparency
Topic Area 2	Risk management and AI systems development lifecy-cle approaches
Topic Area 3	Intended use and analytical and clinical validation
Topic Area 4	Data quality
Topic Area 5	Privacy and data protection
Topic Area 6	Engagement and collaboration

DOCUMENTATION AND TRANSPARENCY

5.1 Documentation and transparency

Documentation and transparency are critical concepts that are essential for facilitating scientific and regulatory assessments of AI systems. They also help ensure trust not only in the AI system itself, but also between developers, manufacturers and end-users. Accurate and comprehensive documentation is essential to allowing a transparent evaluation of AI systems for health. This includes undertaking a total product lifecycle approach to pre-specifying and documenting processes, methods, resources and decisions made in the initial conception, development, training, deployment, validation (data curation or model tuning) and post-deployment of health-related AI systems that may require regulatory oversight. The following discussion focuses on some elements related to documentation and transparency but is not fully inclusive of all of the factors that are relevant to this important area.

Effective documentation and transparency help establish trust and guard against biases and data dredging. The same regulatory expectations and standards that ensure the safety and effectiveness of regulated products also apply to AI systems used in regulated areas. It is important for regulators to be able to trace back the development process and to have appropriate documentation of essential steps and decision points. For instance, aspects requiring careful documentation include specifying the problem that developers are attempting to address, the context in which the AI system is proposed to function, and the selection, curation and processing of training datasets used in the development process.

Documentation should allow for the tracking, recording and retention of records of essential steps and decisions, including justifications and reasoning for deviating from pre-specified plans. Effective documentation may also help to show that developers and manufacturers are taking into consideration the full complexity of the context within which the AI system is expected to operate. Moreover, developers and manufacturers should describe how the AI system is addressing the needs of users and why widening the user base would be appropriate. Without transparent documentation, it becomes hard to understand whether the proposed approaches will generalize from the retrospective clinical evidence presented in the regulatory submission to real-world deployments in new settings, which may markedly reduce performance *(11).* Figure 2 shows examples of essential steps and decision points that developers and manufacturers are encouraged to consider for documentation purposes.

Different entities with multidisciplinary expertise are likely to be involved in the development of AI systems for health and therapeutic development. There is a need to develop a shared understanding of procedures required for transparent documentation and to show that decisions are scientifically sound. Systems used to track and document the development processes and key decision points should record access and should be protected against data manipulation and adversarial attacks.

Documentation and transparency should not be seen as a burden but as an opportunity to show the strength of a science-based development that considers the full context in which the AI system is expected to be utilized, including the characteristics of end-users. Tools and processes for documentation should be proportional to the risks involved. Conversation with regulatory authorities prior to or in the early stages of development is encouraged and may provide vital help in informing documentation needs.

Beyond the regulatory perspective, it is important to note that effective documentation and other steps that help ensure transparency are important ways to establish trust and a shared understanding of AI systems in general. Steps to facilitate transparency include: publishing in peer-reviewed journals; sharing data and datasets; and making code available to foster mutual learning and facilitate additional studies. Academic institutions, medical journals, regulatory organizations and other stakeholders are working on advancing transparency for the use of AI in diagnostic and therapeutic development.

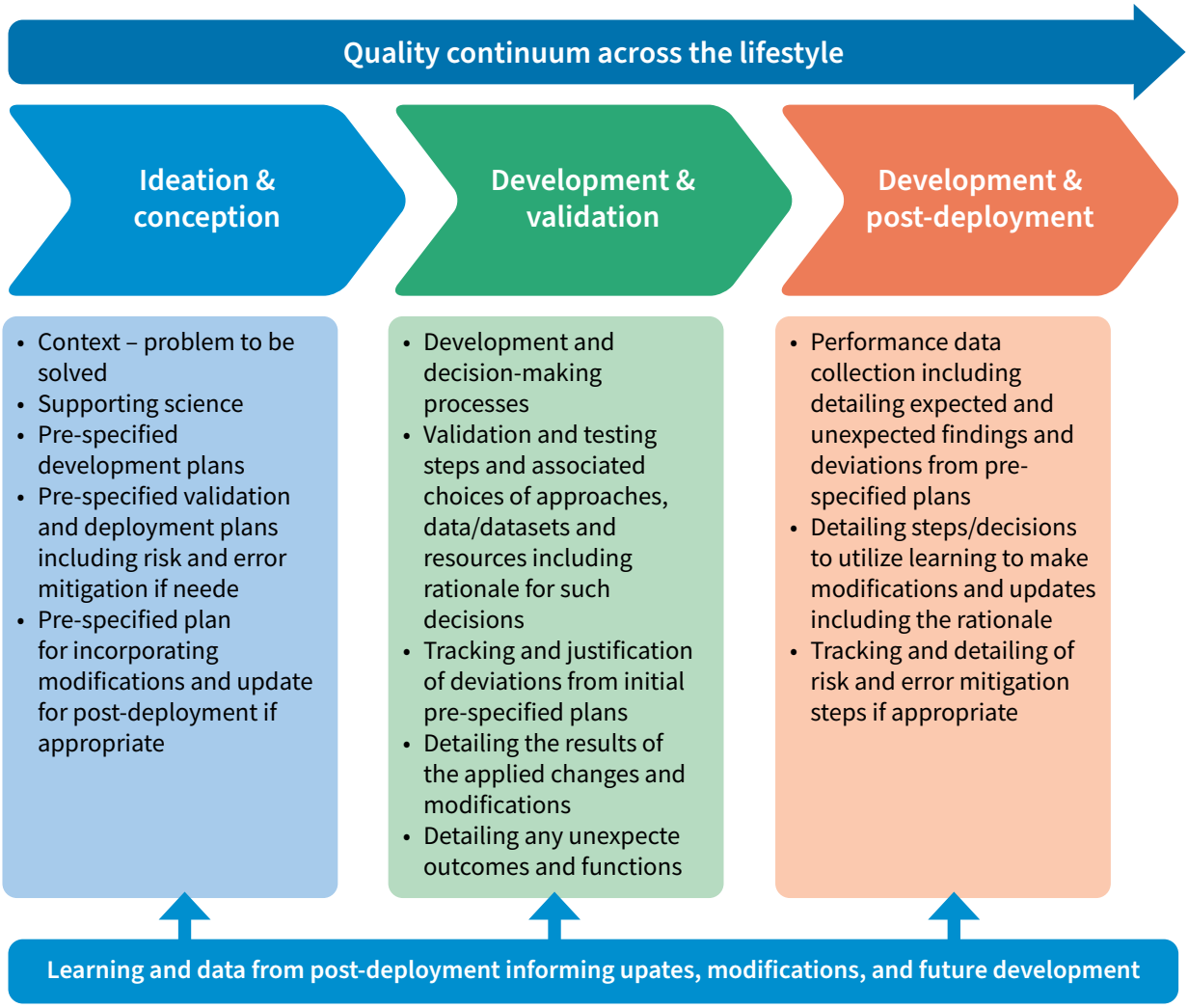


FIGURE 2. Examples of key development decision points in the development of AI systems

Collaborations – such as Consolidated Standards of Reporting Trials for AI (CONSORT-AI) *(12)* and Standard Protocol Items: Recommendations for Interventional Trials for AI (SPIRIT-AI) *(13)* – have given useful guidance about how to design studies to collect clinical evidence where AI systems are used, as well as how to publish the results. Transparency is not only an important consideration for building trust but can also be a useful tool for educating end-users. This can be achieved, if appropriate, by adapting communication strategies to the needs of end-users and other stakeholders such as patients and communities. As outlined in Figure 2, the development process of an AI system is multifaceted. A methodical approach to documentation throughout the full development cycle, including deployment and post-deployment, should be considered.

The following are some elements that might be useful to consider in terms of documentation and record retention.

5.1.1 Documentation across the total product lifecycle – ensuring a quality continuum

Developers should design, implement and document approaches and methods to ensure a quality continuum across the development phases. Effective documentation outlining all phases of development would further enhance confidence in the AI system and would show how expected and unexpected challenges are identified

and managed. Validation processes and benchmarking should be carefully documented – including the decisions for selecting specific datasets, reference standards, parameters and metrics to justify such processes. For example, careful consideration should be given to documenting how and why specific data or datasets are selected to train, externally validate and retrain the model (e.g. post-deployment retraining).

5.1.2 Pre-specification and documenting the medical purpose, clinical context and development

The intended medical purpose/function of the AI systems should be clearly documented. For instance, what is the problem that the AI system aims to resolve? This should take into consideration the full clinical and health contexts in which a tool is expected to function. For example, clinical care environments can be vastly complex and may involve several individuals with different roles and expectations. Documenting how the AI system should function in such active environments must be considered. As shown in Figure 3, there are multiple processes, testing/validation steps and protocols that should be pre-specified and documented. Pre-specification is one of the most important elements that supports trust and confidence in the development process. This will show evidence of a coherent development process and will be the basis for justifying any future changes.

5.1.3 Deployment and post-deployment

AI systems may be designed using data and datasets from specific populations. As with any therapeutics, once deployed, the AI systems will be utilized by a larger population and potentially variable end-users. Careful deployment plans and justification for targeting different end-users should be considered. Manufacturers should be obliged to carry out post-market surveillance, which is the systematic process for collecting and analysing experience gained from AI systems that are considered to be medical devices that have been placed on the market *(14).* Deviations from pre-specified plans, updates or changes to the AI system, post-deployment performance, data capture and approaches to continued assessment of the system should also be documented. Such approaches will be increasingly relevant once there is a wider understanding that AI systems may change after deployment.

5.1.4 Risk-based approach and proportionality

Regulatory frameworks recommend a risk-based approach with processes in place to identify and mitigate errors, biases and other risks in a manner proportional to their importance. A risk-proportional approach should also be considered for the level of documentation and record-keeping for AI systems. Developers of AI systems should keep in mind that regulatory organizations have avenues for dialogue and discussion that can be used to shed light on regulatory requirements.

RISK MANAGEMENT AND ARTIFICIAL INTELLIGENCE SYSTEMS DEVELOPMENT LIFECYCLE APPROACH

5.2 Risk management and artificial intelligence systems development lifecycle approac

AI systems fall into many categories – e.g. devices that rely on AI and are used as medical devices (commonly known as SaMDs, which is short for "Software as a Medical Device"). Such categories of AI systems are defined by the IMDRF as *"software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device" (15).* However, the regulatory considerations for such a category of AI systems are similar to those of typical software that are regulated as medical devices, with the addition of considerations that may include continuous learning capabilities, the level of human intervention, training of models, and retraining *(15)*. Furthermore, a holistic risk management approach that includes addressing risks associated with cybersecurity threats to an AI system, and the system's vulnerabilities, should be considered throughout the total product lifecycle. This topic area aims to present a holistic risk-based approach to AI systems in general, and to those used as medical devices in particular, throughout their lifecycle, including during pre- and post-market deployment.

5.2.1 AI systems during the development and deployment process

Figure 3 illustrates the process of development and deployment of an AI system. Developers and implementers should establish measures to ensure responsible development of AI systems.

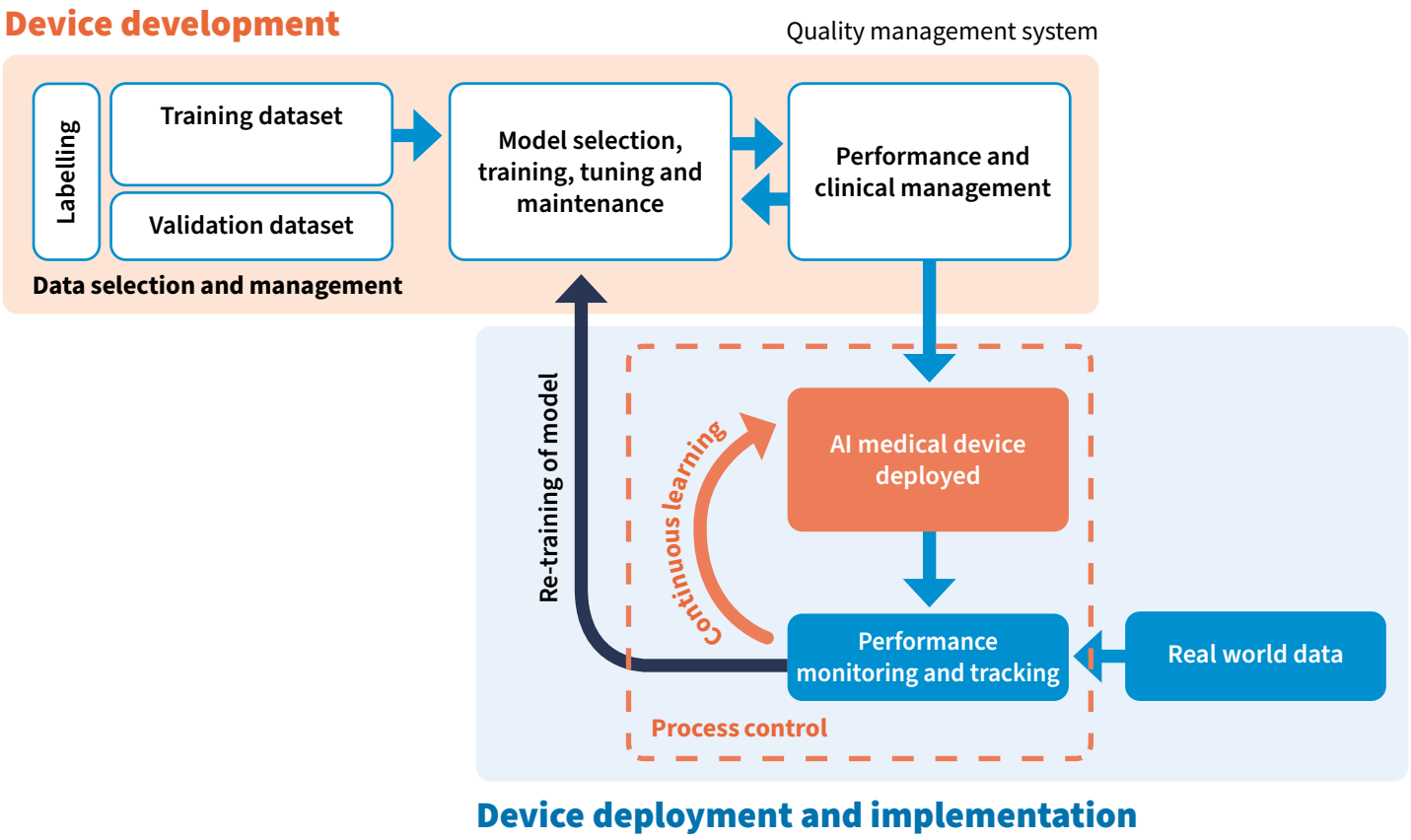


FIGURE 3. The process of developing and deployment of the AI system *(16)*

Figure 3 shows that all activities related to the design, development, training, validation, retraining and deployment of AI systems should be performed and managed under a quality management system based on ISO 13485 *(16)*. For clinical endpoints, AI-specific monitoring dimensions include confidence *(17)*, bias and robustness *(18)*.

5.2.2 AI systems development lifecycle

An AI system development lifecycle approach can facilitate continuous AI learning and product improvement while providing effective safeguards. This can be achieved if the development lifecycle approach involves appropriate development practices for the AI system. This approach could also potentially increase the trustworthiness, and assure performance and safety, of the AI system. An example is the Total Product Lifecycle (TPLC) approach (4) that could include the following four components (as illustrated in Figure 4):

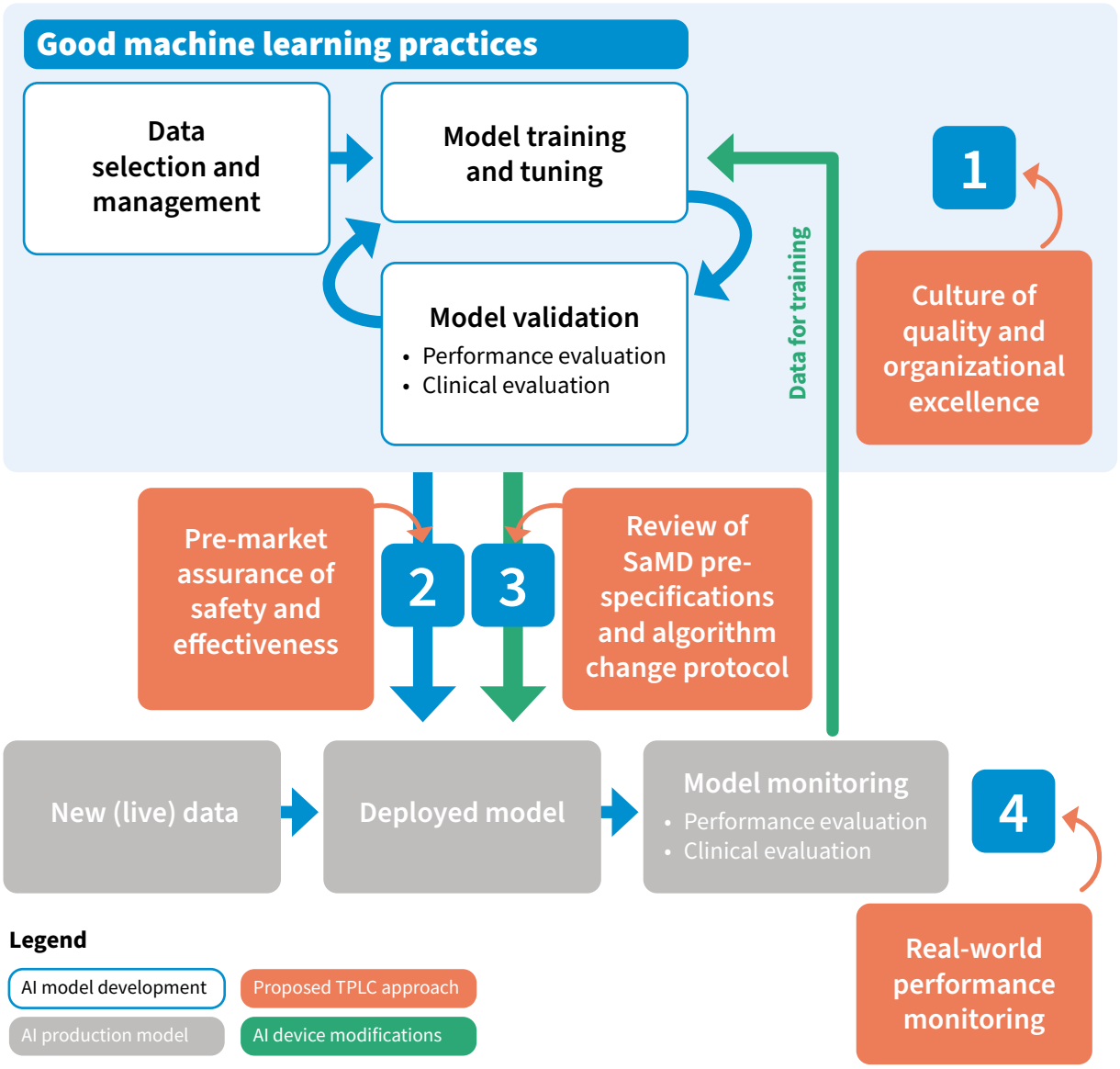


FIGURE 4. AI system Total Product Lifecycle approach on AI workflow *(4)*

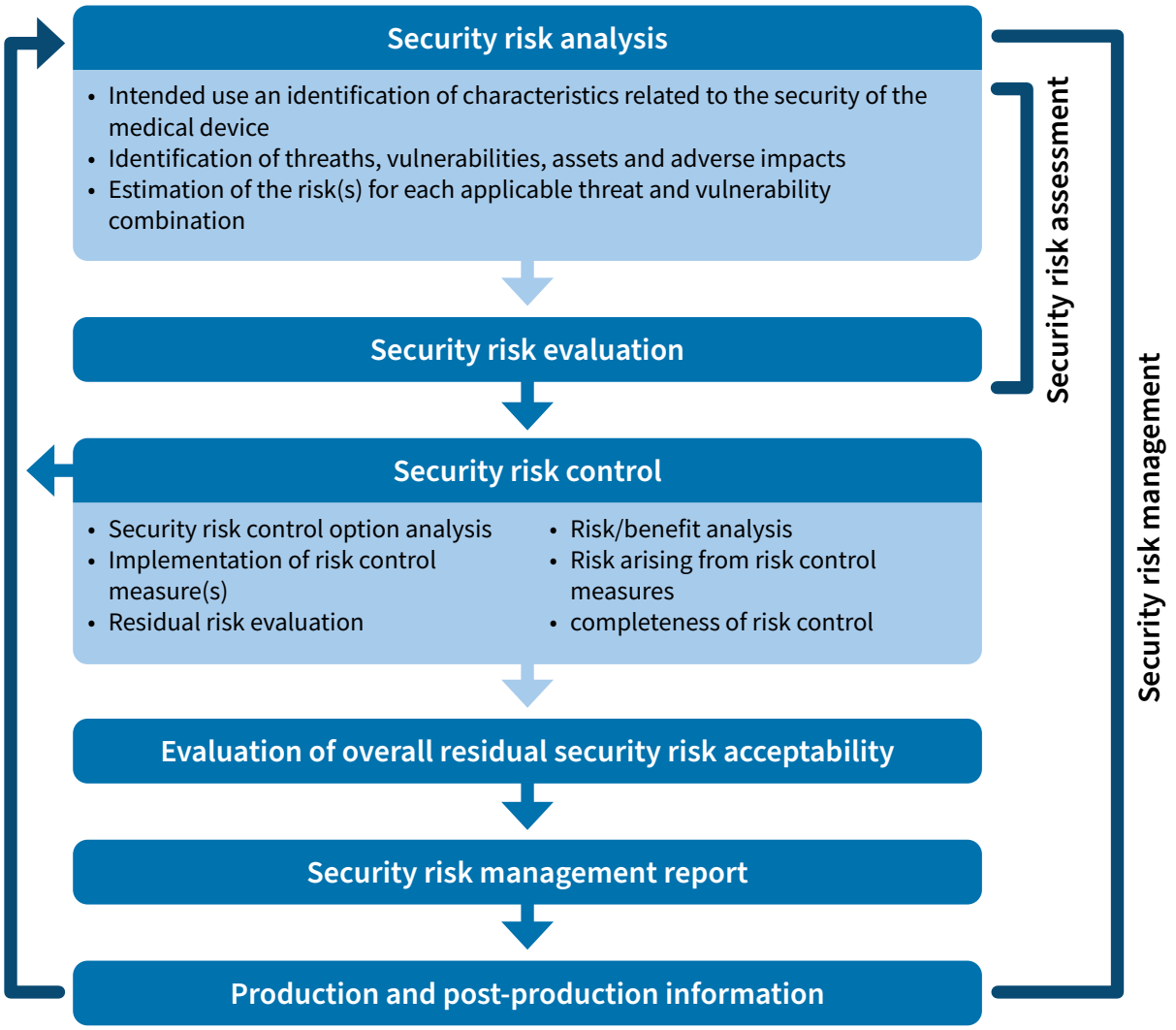


FIGURE 5. IMDRF schematic representation of the security risk management process *(19)*

- demonstration of a culture of quality and organizational excellence of the manufacturer of the AI systems;
- pre-market assurance of safety and performance;
- review of AI systems' pre-specifications and algorithm change protocol; and
- real-world performance monitoring.

5.2.3 Holistic risk management

Holistic risk evaluation and management should be considered, taking account of the full context in which the AI system may be used. This could include not only the software or AI system that is being developed, but also other software that may be used within the same environment or context. Other risks, such as those associated with cybersecurity threats and vulnerabilities should be considered throughout all phases in the life of a medical device. Consequently, manufacturers of AI systems should employ a risk-based approach to ensure that the design and development of AI systems used as medical devices include appropriate cybersecurity protections. Doing so necessitates that manufacturers take a holistic approach to the cybersecurity of the



FIGURE 6. General AI medical device risk management approach

device by assessing risks and mitigations throughout the AI system's development life cycle. In order to achieve this, the IMDRF has published a security risk management process, as illustrated in Figure 5.

However, to facilitate AI systems risk management, a general holistic management approach is introduced in this subsection with three broad management categories: pre-market development management, post-market management and change management. These categories are illustrated in Figure 6 and are discussed below:

- **Pre-market development management**

There is a need for transparency regarding the functioning of any manufactured AI-based devices to ensure that users can have a better understanding of the benefits, risks and limitations of these AI-based systems *(20)*. In addition, the controls and measures put in place to ensure that a developed AI system functions as expected while minimizing risk of harm should be proportional to the risks that could occur if the AI system were to malfunction. For instance, failure of an AI system that is designed to encourage adherence to a healthy diet is different from one that is designed to diagnose or treat certain diseases and pathologies. Therefore, developers should consider a risk-based approach through all processes to prioritize safety. Developers need to consider both the intended use of the AI system and the clinical context in order to evaluate the level of risk. For instance, the IMDRF risk framework for SaMD *(21)* identifies two major factors that may contribute to the impact or risk of an AI system. The first factor is the significance of the information provided by the AI system to the health-care decision. The significance is determined by the intended use of the information – to treat or diagnose, to drive clinical management, or to inform clinical management. The second factor is the patient's health-care situation or condition – which is determined by the intended user, disease or condition, and the intended population for the AI system – i.e. critical, serious or non-serious health-care situations or conditions. Taken together, these factors relating to the intended use can be used to place the AI system into one of four categories from lowest risk (I) to highest risk (IV) to reflect the risk associated with the clinical situation and device use.

The intended use and risk classification should be considered when testing different models and balancing trade-offs such as transparency and accuracy. In cases where training datasets are limited, simpler models, such as regression or decision-tree models, often provide equivalent or better results than more complex models and have the added benefit of more transparency and interpretability. On the other hand, in cases with larger and more complex datasets, complex models such as deep learning networks may not lend themselves to being explainable but may provide greater accuracy than simpler models. However, in cases in which there is a greater risk of harm, stakeholders should consider discussing the risks and benefits of choosing a more complex model and whether there are ways to mitigate the lack of interpretability and transparency and to build trust in the model through additional validation measures.

Table 2. AI systems risk classification *(21)*

State of health-care situation or condition	Significance of information provided by the AI system to the health-care decision		
	Treat or diagnose	**Drive clinical management**	**Inform clinical management**
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Furthermore, depending on the level of risk, some AI systems may be approved as being available for full deployment whereas others may be initially authorized for deployment in more "AI-ready" institutions. "AI-ready" institutions are those which are certified on the basis of having stringent levels of surveillance in place with responsive back-up systems to handle any failure of the algorithm in order to minimize risk of patient harm.

Overall, it is important to achieve transparency between all AI-system stakeholders, including the developers, manufacturers, regulatory authorities and implementers (i.e. users in health-care settings, such as medical practitioners). Appropriate documentation of risk management and proper auditing procedures are examples of ways that help assure transparency. In general, auditing of specific key components of the AI medical device should be considered (e.g. certain software, hardware, training data, failure cases). For instance, it is important to do version control with training data because more data are added with each update. If an algorithm suddenly deteriorates in performance after an update, an inspection of everything that contributed to the update may be desired. In most cases, the element that will have changed is the addition of new training data by the developer (rather than changes to the software itself, such as modification to the neural networks). Moreover, given how unpredictable changes in performance can be for AI, it is recommended to have active reporting and investigation of failure cases (in the CONSORT-AI guidelines) – although it is not prescriptive, given the wide range of available reporting and investigation avenues from common-sense clinical auditing (i.e. human inspection) to technical solutions based on inference.

Although not specific to AI, there is a thickening web of country-, nation- and jurisdictional-specific legislations and laws that manufacturers and developers may need to consider for the development and deployment of regulated AI medical devices in health care. Such legislation includes the Personal Data Protection Act, Human Biomedical Research Act, Private Hospitals and Medical Clinics Act, Health Insurance Portability and Accountability Act and General Data Protection Regulation (GDPR). Compliance with relevant laws (local, cross-jurisdictional laws and data protection acts) needs to be demonstrated by manufacturers and developers of medical devices whether they embed an AI component or not.

- **Post-market management**

Post-market monitoring and surveillance of AI medical devices allows timely identification of software- and hardware-related problems which may not be observed during the development, validation and clinical evaluation of the device. New risks may surface when the software is implemented in a broader real-world context and is used by a diverse spectrum of users with different expertise. Companies involved in distributing AI medical devices (manufacturers, importers, wholesalers, authorized representatives and registrants) are required to comply with their post-market duties and obligations which include reporting to relevant regulatory authorities in any of the following circumstances *(14,16)*:

- any serious public health threat;
- death, serious deterioration in the state of health of patient, user or another person has occurred;
- death, serious deterioration in the state of health of patient, user or another person may have occurred;
- any field safety corrective action (such as return of a type of device to the manufacturer or its representative [also known as recall in some jurisdictions]; device modification; device exchange; device destruction; advice given by the manufacturer regarding the use of the device).

Furthermore, manufacturers should proactively collect information (through scientific literature and other information sources such as publicly accessible databases of regulatory authorities, user training and surveys) as part of their post-market surveillance plan. The plan should outline how manufacturers will actively monitor and respond to evolving and newly-identified risks. Key considerations for the post-market surveillance plan include *(16)*: vulnerability disclosure, patching and updates, recovery and information-sharing. Additionally, as part of the post-market duties and obligations, companies involved in distributing medical devices (manufacturers, importers, wholesalers and registrants) are required to report adverse events associated with the use of software medical devices to relevant regulators.

In general there is a need for both post-market clinical performance follow-up and periodical safety checks to report any potential harm. The intensity of post-market surveillance by the manufacturer may be risk-proportionate (according to consequences of failure [creating potential risk of harm] and likelihood of early detection of such failure). Finally, post-market surveillance requires a minimum level of evaluation for each site in order to ensure that potential algorithm vulnerabilities due to variation in local environments can be detected.

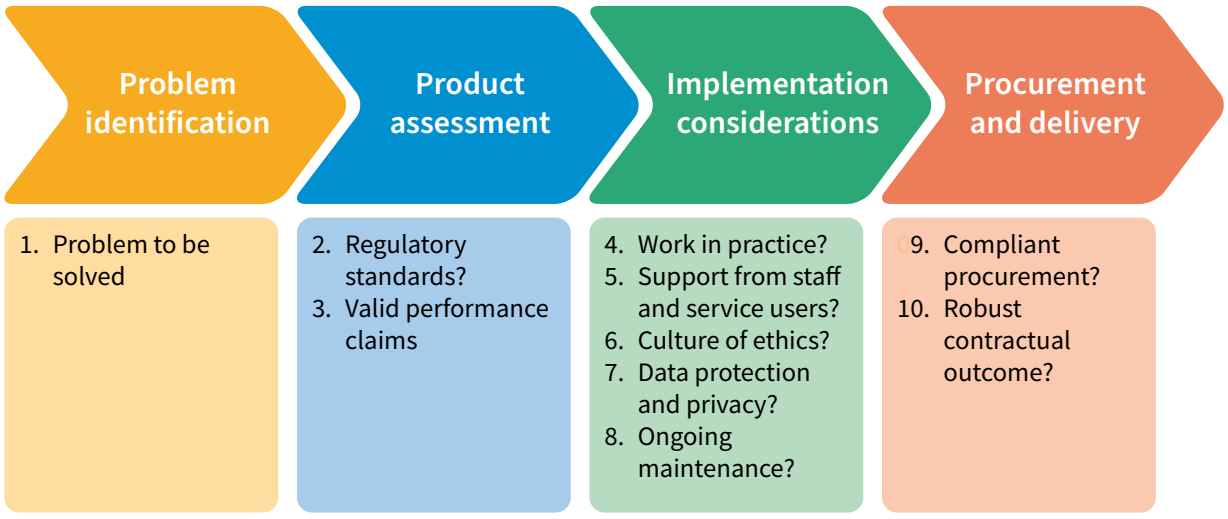


FIGURE 7. The United Kingdom's National Health Service *A buyer's guide to AI in health and care (22)*

For example, the AI Lab of the National Health Service (NHS) in the United Kingdom of Great Britain and Northern Ireland published guidance to accelerate a safe and effective adoption of AI in health *(22)*. The guide lists 10 questions in four categories to help buyers of AI products to make informed decisions, identify problems, assess products, and consider issues relating to implementation, procurement and delivery (Figure 7).

- **Change management**

In view of the character of AI systems, it is important that the regulatory system enables continuous modifications for improvement to be made throughout the AI system's development lifecycle. The term "change" refers to such modifications, including those performed during maintenance.

There are several proposed change management models and approaches for AI-based systems. Some consider change as part of the total development lifecycle (as in the TPLC approach) *(4)* (Figure 4). Other models focus on the change management process in the total lifecycle of medical device products which can be continuously improved. An example of this is the approach implemented by the Ministry of Health, Labour and Welfare of Japan and adapted in the Pharmaceuticals and Medical Devices Act as Post-Approval Change Management Protocol (PACMP) for medical devices *(23)* (Figure 8).

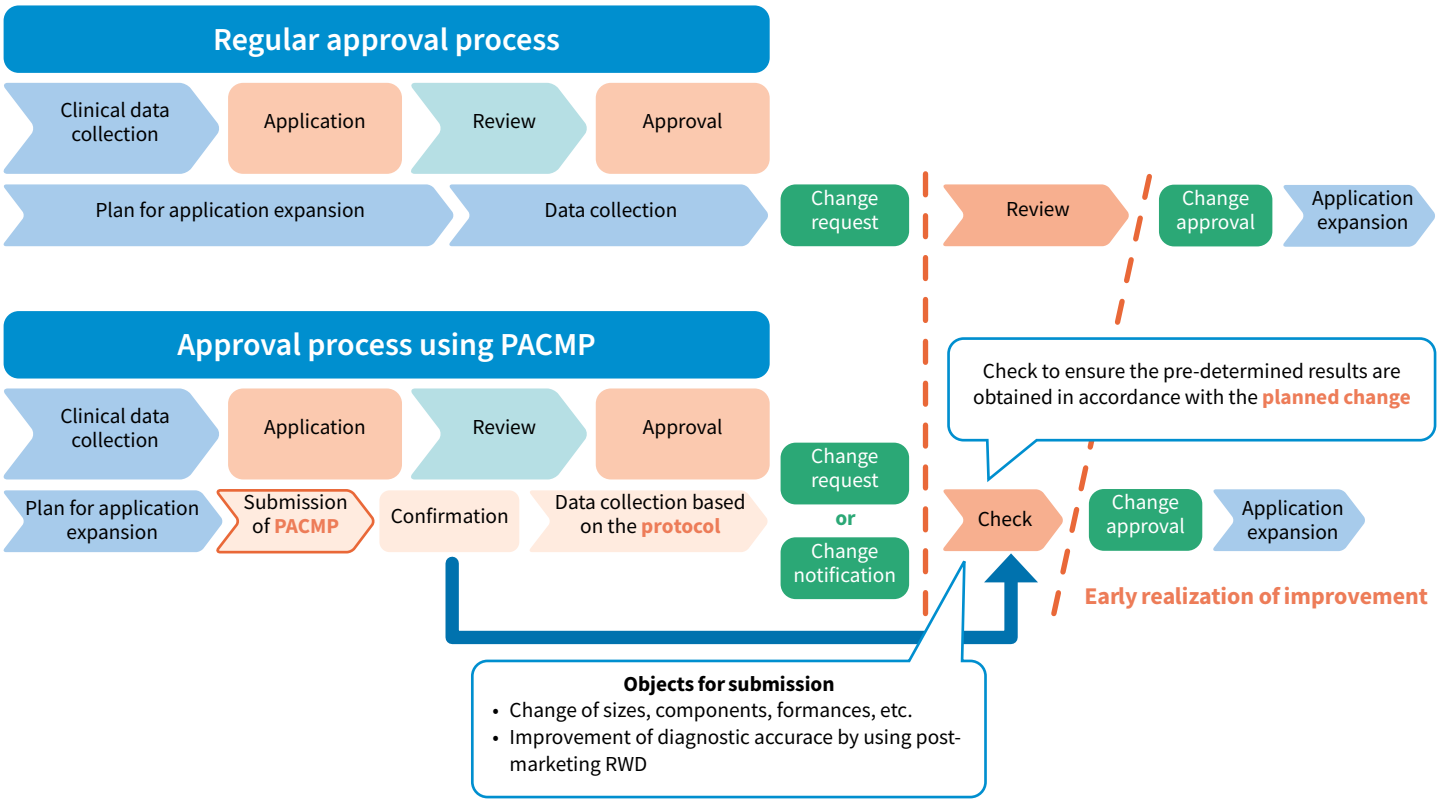


FIGURE 8. Post-Approval Change Management Protocol for medical devices

INTENDED USE AND ANALYTICAL AND CLINICAL VALIDATION

5.3 Intended use and analytical and clinical validation

In principle, regulatory mechanisms are in place to answer the question: "Do the available data (included in the regulatory submission) support the conclusion that an investigational or experimental AI system is safe and performs sufficiently well to justify entry into the market and public access?" In addition to the principles discussed in 5.1 and 5.2, one also must consider assessing if the use of the system is safe (i.e. it will not harm the user, the patient or other persons) and if the claims made about its performance can be verified (see Figures 9 and 10). Evaluation of these claims for AI systems requires a clear use case description, demonstration of analytical and clinical validation, and assessment of the potential for bias or discrimination in the AI system.

5.3.1 Use case description, analytical and clinical validation

Clinical evaluation is the review of evidence that demonstrates the safety and performance of a given product for a given intended use. For AI systems (especially devices that rely on AI and are used for medical purposes), guidance is useful for collecting evidence of analytical and clinical validation. The performance of AI systems can be changed rapidly – not only as a result of a code change but also to provide different or additional training/tuning data. Consequently, clinical evaluation that takes account of TPLC from development to analytical and clinical validation and to post-market surveillance should be considered for AI systems.



FIGURE 9. Domains of health technology regulation, assessment and management for drugs and devicesh

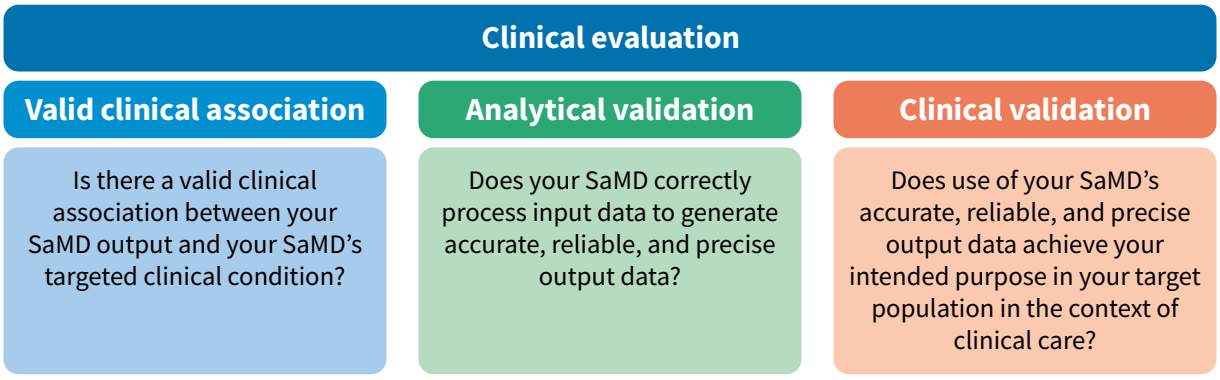


FIGURE 10. IMDRF description of clinical evaluation components *(4)*

This topic area covers the considerations of use case descriptions (including statements of intended use) and analytical and clinical validation. These considerations follow the framework proposed by the WHO/ITU FG-AI4H Working Group on Clinical Evaluation (WG-CE) (24). A full description of this framework can be found in the deliverable for the WG-CE. The following section describes the key considerations and best practices, and builds on the important work of national and regional regulatory authorities and international bodies such as IMDRF. It is not intended to replace the work of these bodies. By outlining key considerations, this report draws attention to challenges that remain in this rapidly changing field. For instance, particular consideration is given to under-resourced settings which may have limited regulatory capacity at national level. The role of benchmarking in the evaluation of AI systems in health is also explored. Evaluation principles are applied to this topic area, and to the work of the WHO/ITU FG-AI4H in which benchmarking evaluation is a key component *(25).*

5.3.2 Intended use

AI systems are complex, dependent not only on the constituent code but also on the training data, clinical setting and user interaction. They are often situated in a complex clinical pathway or are being introduced into new clinical pathways altogether (e.g. into new telemedical pathways or as part of new triage tools). Therefore, for AI systems, safety and performance can be highly context-dependent. The description of the use case has a substantial role both to inform end-users where the tool can be utilized safely and appropriately and, in regulated AI systems (the statement of intended use), to allow regulators to assess whether the evidence of the analytical and clinical validation steps is appropriate and sufficient for the intended use.

When developing a health-related AI system, it is important to describe the relevant use case. This consideration should cover the setting (geography, type of care facility), the population (ethnicity, race, gender, age, disease type, disease severity, co-morbidities) the intended user (health-care provider or patient) and the clinical situation for which it is intended. Many interventions, tests and guidelines are prone to bias, and this is a particularly important consideration for AI systems which are highly sensitive to the characteristics of the data they were trained on and are prone to failure with unseen data types (such as a new disease feature or population type or context that was not previously encountered). Developers and manufacturers should also provide a clear clinical and scientific explanation of their tool's intended performance, including the populations and contexts for which it has been validated for use. Standardized reporting templates common to all stakeholders can help to communicate the intended use more effectively *(26, 27, 28)*. For some intended use cases there may be clear reasons why analytical performance of the tool would differ in different settings *(29)* (e.g. a symptom checker may perform differently in areas with a disease epidemiology that is different from the data on which it was trained). If this is the case, systematic known differences in performance should be included in the intended use statement. For other intended use cases, there may be emerging evidence that the tool under consideration, or another very similar tool, has been shown to have similar analytical performance in a wider setting than those in which the tool was initially developed and validated (30) (e.g. retinal tools have been shown to have a similar performance in different populations *(31)*). Understanding of the generalizability of similar tools may be taken into account when providing a statement of the intended use or description of the use case *(32).*

As part of the risk management process, regulators may wish to request evidence that developers have considered whether there are situations in which a tool should not be used (e.g. if there are insufficient training data for a particular patient group, or absence of validation in a particular setting), or if there are potential risks from use outside of the intended settings.

5.3.3 Analytical validation (also referred to as technical validation)

For the purposes of this document, analytical validation refers to the process of validating the AI system using data but without performing interventional or clinical studies. This may also be referred to as technical validation. Appropriate analytical validation demonstrates that a model is robust and performs to an acceptable level in the intended setting. It also enables the understanding of potential bias and generalizability (and any steps taken to understand these).

Developers and manufacturers should provide a description of the datasets used in the AI system's training, tuning, testing and internal validation. The description of the intended use case (which can be on standardized reporting templates) should cover the size, setting, population demographics, intended user and clinical situation (with input and output data). Transparency and documentation on dataset selection and characteristics are critical to ensure that AI systems are used appropriately. Developers and regulators may expect that the AI system has been externally validated in a dataset different from that in which it was trained and tested in order to demonstrate the model's external validity and generalizability beyond the original dataset. The external validation dataset is expected to be representative of the setting and population that are described in the intended use (gender, race, ethnicity) in order to demonstrate robust performance in the intended setting. The validation dataset should be of adequate quality, with appropriate robustness of labels. As part of the risk management process, it is important to identify any cases that are, or may be, high-risk *(28).*

Although bias, errors and missing data are not unique to AI development, they are nevertheless serious concerns, which may arise for many reasons – including unequal and non-representative training or validation datasets, or structural bias in the systems where training data is generated (e.g. health-care settings). Reporting the gender, race and ethnicity of persons in the training and validation data cohorts, if feasible, can help to

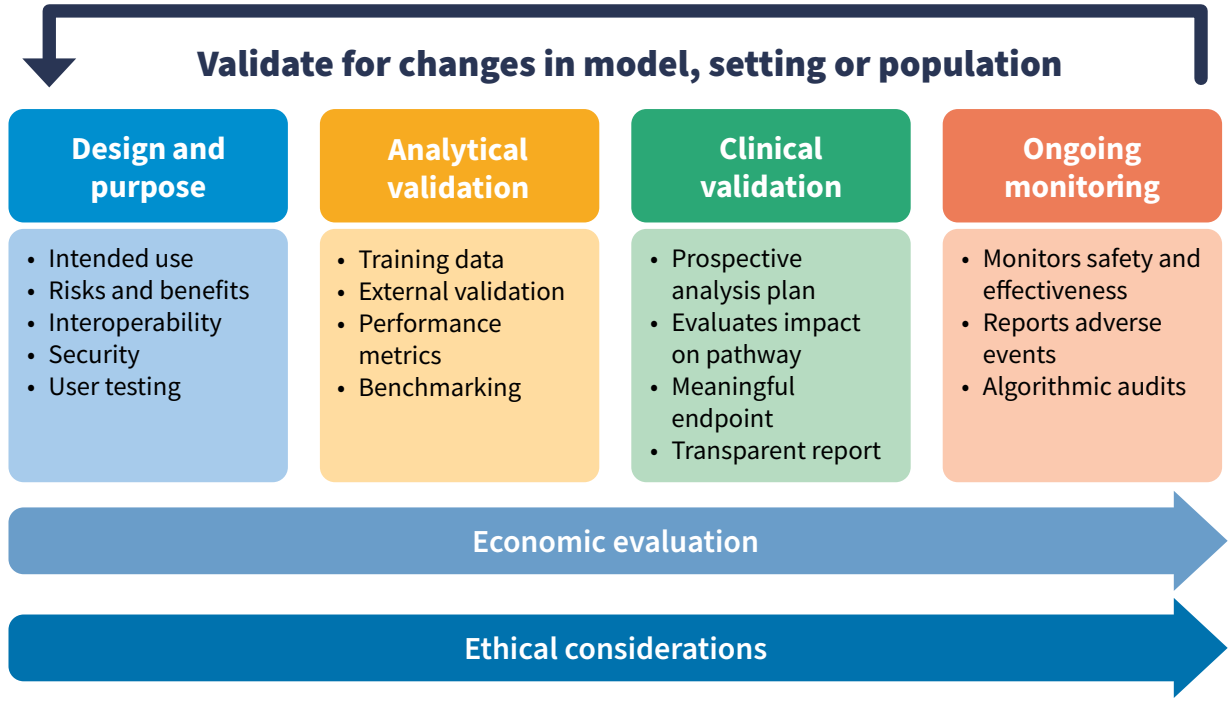


FIGURE 11. Overview of framework for clinical evaluation of AI models in health developed by the WG-Clinical Evaluation

address the potential for bias and can avert its impact. For example, a better understanding of bias may help identify populations for which an AI system may not function as expected. Post-market surveillance can also provide insights into the impact of potential bias.

Obtaining datasets for training, testing and validation that are sufficiently representative and of sufficient quality can be difficult. Local, regional and national bodies interested in procuring AI systems could hold their own hidden dataset to enable external validation (e.g. a recent scheme of the United Kingdom of Great Britain and Northern Ireland's NHSX has nationally-representative datasets for some common use cases). Access to representative datasets for validation is a particular concern in many low- and middle-income countries. Where datasets are available in low-resource settings, there may also be limitations in the quality of the data. The ability to produce robust datasets with high-quality ground truth labels is likely to be affected by limitations elsewhere in the health setting where there may be barriers that impede access to diagnosis and treatment. These major challenges – which have the potential not only to propagate inequality of access but also to compromise safety and performance of AI-based tools – are potential areas for future work. In this regard, the newly launched International Digital Health & AI Research Collaborative (iDAIR) *(33)* notes that collaborative, distributed and responsible use of data is at the heart of its strategic plan.

While most regulatory agencies have national or regional remits, some countries with limited regulatory capacity tend to rely on decisions made by other major regulators. The availability of independent, hidden, representative datasets also offers particular advantages to countries that do not have their own regulatory process, or where regulatory decisions may be informed by dossiers provided to other bodies. However, the performance of AI-based systems is highly dependent on the context. In order to rely on regulatory review and decisions, many regulators (whether national or regional) could perform analytical validation as a second local validation step to ensure that the performance metrics obtained are consistent with those demonstrated in other regulatory jurisdictions. This could be best prioritized through a needs-based approach – e.g. the identification of key areas in which AI-based tools are promising and could provide local value – and the potential prospective creation of datasets to support validation.

In order to understand the performance of an AI system, evaluation against an accepted standard should be made. The most appropriate standard for comparison may differ by intended use but commonly-used standards are human performance in a similar task or other models (e.g. derived from logistic regression) with strong evidence-based or mandated standards of accuracy, sensitivity and specificity (such as for screening tools). Depending on the intended use case, the requirement for comparative performance may be more or less stringent (e.g. when used as a triage or screening tool, a different level of comparative performance may be acceptable compared to a tool used for diagnosis).

Some limited comparative benchmarking of AI systems has been performed in a single setting but may become more common as the number of available tools increases *(34)*. In the future, if an AI system has proven clinical efficacy and safety in a particular setting, it may be possible and appropriate to benchmark other newer tools against that AI system to understand potential similarity of performance. Benchmarking software is being developed as part of the work of the Open Code Initiative *(35).* Platforms such as this may also be useful as ways to perform repeated algorithmic validation of models that have been updated. However, this is currently not the case for any use cases, and benchmarking thus far has been used only to understand comparative analytical performance. In addition, repeatedly using the same data for benchmarking multiple updated models (and thus, even if inadvertently, for training the test) risks introducing bias, and this should be taken into account when benchmarking is considered.

A designated FG-AI4H working group on data and AI solution assessment methods (36) provides guidance on the methods, processes and software development for the analytical validation of health-related AI systems *(28).*

5.3.4 Clinical validation

Analytical validation performed retrospectively on an existing dataset gives measures of performance (accuracy, negative predictive value, positive predictive value) but does not allow for evaluation of other factors that may affect performance of the tool (e.g. user interaction, workflow integration, and unintended consequences of the tool within a complex clinical pathway).

Both national and international bodies have proposed a graded set of requirements based on risk for digital health tools (including significance of the information provided by the tool and the state of the health condition) *(37, 38)*. The IMDRF document on clinical evaluation of SaMD (Table 2 *(21)*) proposes that devices in category I are the lowest-risk tools that have evidence of analytical validity, and that a novel tool in this category would require manufacturers to collect real-world performance data and generate a demonstration of scientific validity. For higher-risk SaMD, clinical evaluation evidence is expected on the basis of evidence of analytical validity. There is no universal agreement on the appropriate level of evidence of adequate clinical performance for a novel AI tool before deployment and this is the subject of a separate working group within the FG-AI4H (Working Group on Clinical Evaluation).

Randomized clinical trial data are the gold standard evaluation of comparative clinical performance, and may be appropriate for the highest-risk devices where an AI tool has no demonstrated performance in that setting, or for large national procurement bodies that seek evaluation of performance before national expenditure. A trial that is expected to guide clinical practice should have a clinically meaningful primary endpoint (morbidity, mortality) but, in certain situations, event rate or time lag between the trial and the endpoint may result in a more feasible surrogate endpoint. Reporting guidelines backed by the widely accepted EQUATOR network are now available for protocols and clinical trials using AI systems *(12)*. However, currently there remain a small number of actively recruiting or completed randomized trials in this field *(39).*
Randomized clinical trials have potential limitations that may make other options preferable (trials can be slow, or expensive, and may evaluate performance in specific groups under trial conditions). Where randomized evidence may not be necessary (e.g. the evidence required may be proportional to the risk or cost of a tool), prospective validation in a real-world deployment and implementation trial, with a relevant comparison group showing improvement in meaningful outcomes using validated tools or widely accepted and verified endpoints and with systematic safety reporting, may be appropriate. Clinical performance should be considered in the context of the capability of the health workers, available Internet bandwidth and health informatics infrastructure, and real-time data pipelines. Developers should provide a description of the steps taken to perform clinical validation in a context similar to that available in the intended use setting.

Further consideration of the most appropriate level or type of clinical evaluation for a digital health intervention will be provided by the WG-CE.

In some situations, as described below, special considerations apply. For instance:

5.3.5 Post-market monitoring

Post-market monitoring in some regulatory contexts relies heavily on reporting of adverse events. Recent WHO guidance recommends that proactive post-market surveillance must be carried out by the manufacturer.

As part of a TPLC approach to regulation in this context, further prospective post-market clinical follow-up should be completed after deployment. Regulators must be notified of reportable incidents (adverse events), and findings from more continuous monitoring using real-world data may help developers and regulators better understand and assure the safety and performance of these devices in real-world use. For prospective monitoring of real-world data, significant investment will be required in prospectively curating and labelling validation data. A defined period of close monitoring may be appropriate for AI-based tools for those with high risk given their tendency to overfit on erroneous data features and produce unpredictable errors on unseen data features combined with the lack of data from use in real-world settings with long-term results. Regulators may recommend that manufacturers develop specific market surveillance measures that are appropriate for AI systems.

5.3.6 Changes to the AI tool

An update of an AI tool by a change of code, change of the user interface or provision of further training data may alter the analytical or clinical performance of an AI system. The group are not aware of currently-approved medical AI systems that are "continuously learning" but anticipate that these may be developed. Such AI systems would require a risk–benefit evaluation in keeping with the concepts in this document and with the clinical evaluation of AI systems for health. Taking "checkpoints" – by evaluating the tool as it is currently performing at regular intervals – enables regular evaluation and could signal changes in performance. Depending on the risk of the AI systems and the extent of the changes, appropriate validation must be agreed by the developer and the regulator. Analytical validation against previously unseen datasets – or benchmarking against approved datasets representative of the intended setting or population – could be useful in this scenario.

5.3.7 Low- and middle-income countries

There is considerable variation in the implementation regulation for medical devices, and therefore also in deployed AI technologies and developed AI systems. Some countries lack a dedicated national regulatory body. The WG-RC meetings have provided a forum for the sharing of expertise and discussion of common problems, including for regulatory bodies and other interested stakeholders, some of whom have aligned remits. Furthermore, there are important regulatory considerations related to the intended use and analytical and clinical validation of AI systems in health. First, in low- and middle-income countries, one of the potential uses of AI technologies is in bringing specialized AI-based systems or knowledge to areas which do not have a relevant medical specialist (e.g. interpreting retinal scans, histopathology slides or radiology images). In high-income countries, AI systems are more often positioned as an adjunct to medical professionals. Using an evaluation performed to support regulation in a high-income setting to inform how such AI systems are used in low- or middle-income settings may therefore not be appropriate. Thus, the full context of health-care infrastructure and resources should be considered. Second, some regulatory bodies rely on decisions from other bodies to support their regulatory work. Given that the performance of AI systems may be highly context-dependent, additional steps may be required. There is a concern that developers may not ensure adaptation or evaluation for resource-limited settings if the market there is less attractive. Regulatory agencies in high-income countries could support this adaptation, which could also increase the generalizability and robustness of AI systems. However, this would require adaptive studies to ensure wider use in low- and middle-income countries or the use of incentives to encourage additional development, testing and validation. The availability of a range of representative datasets would support local analytical validation. Finally, AI systems for health can be highly sensitive to shifts in data distribution and features. They may therefore be sensitive to differences in disease prevalence when moving from high-income to low-income counties, with the possibility of lower performance without appropriate evaluation or tuning with local data.

DATA QUALITY

5.4 Data quality

5.4.1 Data in current health ecosystems

The health sector has been very receptive to the benefits of AI thanks to the explosion of data and accessibility to computational power. Data are the most important ingredient for training AI/ML algorithms, and can be classified on the basis of format, structure, volume and many other factors. Data can take any form, including character, text, words, numbers, pictures, sound or video. Also, these data can be structured, semi-structured or unstructured *(9).* Structured data are normally stored in databases that are structured in a manner that follows a specific model or scheme – such as data stored in electronic medical records, mobile devices and Internet of Things (IoT) devices. Regardless of the format, structure or volume of the data, a more general classification can be based on the following 10 Vs of data *(9)* (as illustrated in Figure 12): Volume, Veracity, Validity, Vocabulary, Velocity, Vagueness, Variability, Venue, Variety and Value.

5.4.2 Good quality data in health AI systems

All AI tasks and solutions use some form of data, regardless of their characteristics, to facilitate machines to learn, adapt and improve on their learning. However, data quality greatly influences the success of such solutions' safety and effectiveness. "Good-quality data" is an ambiguous term that is open to misinterpretation. Therefore, gaining a good understanding of the datasets used, for example, from the 10 Vs perspective is crucial to assess data quality in AI systems during development and even afterwards. Section 5.4.3 highlights key challenges and considerations for all stakeholders, including developers and regulators, when handling data in AI systems in order to achieve good data quality.

5.4.3 Key quality data challenges and considerations for health AI systems

The availability of good-quality datasets that are clinically relevant is one of the key challenges that developers face. However, data of varying quality can still be used depending on the purpose, and thus developers should determine if available data are of sufficient quality to support the development of systems that can achieve their intended goal. The lack of good-quality datasets for use in the development of AI systems may hinder their effectiveness and potential benefits. Data that are not of sufficient quality for the intended purpose can also lead to many problems, such as bias and errors. Some data quality issues that often arise when developing AI systems, and that need to be considered by all stakeholders, are discussed in this section and summarized in Table 3. These issues and considerations can relate directly to dataset management, the ML model, the infrastructure used to manage the data, or general governance aspects, as follows:

- **Dataset management.** When managing datasets for ML models, a clear data management plan should be pre-specified and well documented. Data management approaches should be risk-based and fit for purpose. This may include data selection volume (including volume of data used and volume of available data), splitting, cleansing (including any AI algorithms that were used to clean the data), data usability (including how well the dataset is structured in a machine-readable format), labelling, dependencies, augmentation and streaming. If data augmentation is relevant, it is important to develop a clear data augmentation strategy. The developers should also consider putting in place good data accountability practices for those handling the data in order to ensure that data quality and integrity are maintained throughout the lineage of the data. This is also essential for knowledge management and transfer in a highly evolving field. Further, in addition to the handling of the data, the capacity to plan for and conduct data analyses is also important.

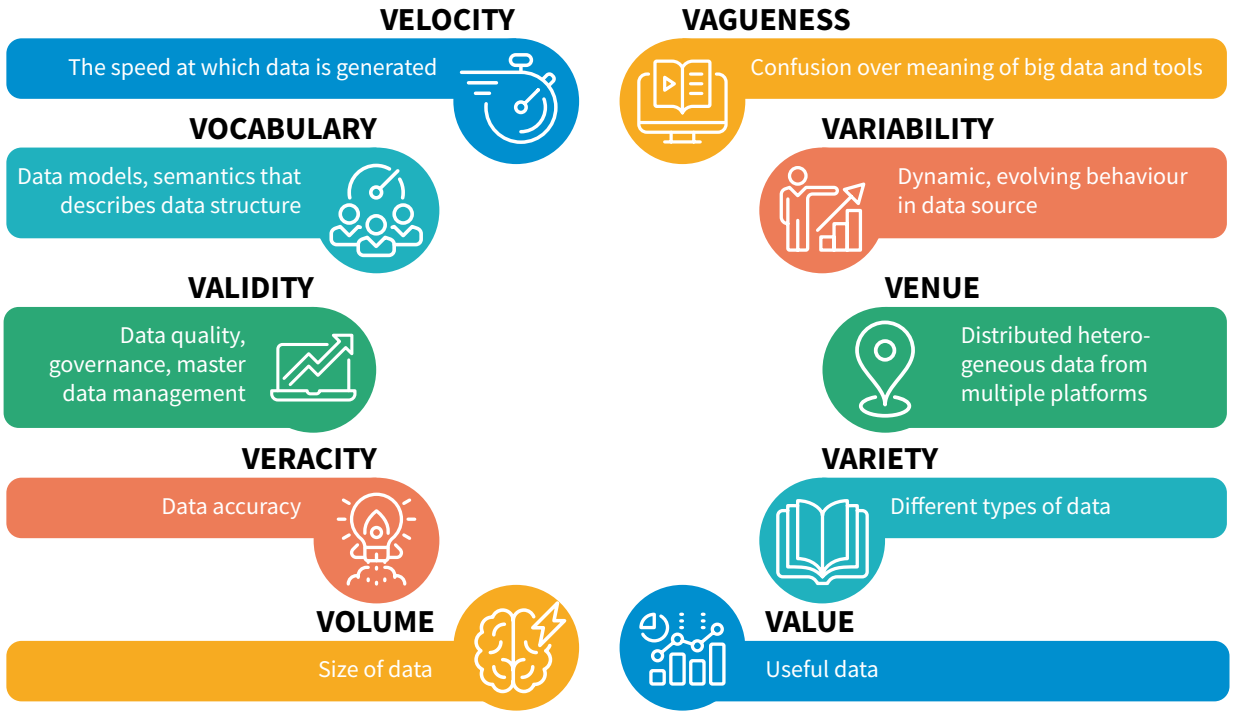


FIGURE 12. The 10 Vs of data *(9)*

- **Data inconsistency.** High heterogeneity in the syntax of the data may require harmonization in order to address issues related to multiple data sources with varying standards, formats, schemas, structures and ambiguous semantics and generate a coherent dataset for the purpose comprehensive analysis – which is especially challenging when using health-care data. For instance, much of the data collected from various information silos is inconsistent, incompatible or not executable in machine-readable formats. For multiple data sources, there may be variations in how the data are captured (e.g. definitions of individual variables).
- **Dataset selection and curation.** Knowing the source of data and making an initial assessment of the data quality can help to determine the potential for selection and information bias. Selection bias results when the data used to produce the model are not fully representative of the actual data that the model may receive or of the environment in which the model will function. In addition to selection bias, measurement bias is another relevant issue that results when the data collection device causes the data to be systematically skewed in a particular direction. Consequently, developers should be aware of data quality limitations when attempting to curate and utilize these large-scale datasets. Moreover, developers and regulators need to know where the data originally came from and how the information was collected and curated. This is especially important when the datasets are from an open-source database where the original source and specifications of the dataset may not be available. When the origin of data is difficult to establish, it would be prudent for developers to assess the risks of using such data and manage them accordingly. Finally, even if datasets are collected from reliable sources, the mitigation of bias and assessment and mitigation of other risks to data robustness remain essential for a heterogeneous dataset.

- **Data usability.** It is essential to know whether the data used for development of the algorithm was intended for that training, so developers need to convey their full understanding of the dataset and why it was suitable for their purpose. For instance, data from a third-party source may be representative data intended for training purposes (e.g. case studies in tertiary education) and may not be suitable for training an AI model intended to diagnose a disease or condition.

- **Data integrity.** Data integrity can be defined as "the completeness, consistency, and accuracy of data" *(40).* Lack of data integrity is an important issue. This can be best understood by how well extraction and transformation have been performed on the dataset. To maintain data integrity, data verification checks may be developed. Data verification checks are a key component of data quality assurance when utilizing real-world data. Such checks should also be the first step in data preparation for any ML workflow.

- **Model training.** AI algorithms are usually trained on a separate dataset (called the training dataset) and validated on a different dataset in order to measure the performance of the algorithm reliably. Training datasets should be well represented (e.g. by considering the prevalence of a disease/condition) to avoid "class imbalance". Medical record data is inherently biased, and therefore it is necessary to incorporate non-medical data such as the social determinants of health *(42).* Furthermore, under-representation of important diagnostic features may limit the performance of the model and cause bias. This can be avoided by ensuring that inclusion and exclusion criteria at the patient level and the data input level do not create a selection bias. Furthermore, when ensuring that the datasets reflect the setting in which the model will be applied, a lack of diverse data (age, race, geographical areas) could limit the generalizability and accuracy of a developed AI system. This is demonstrated by a recent study by Stanford University *(43)* which showed that 71% of patient data from just three US states train most of the AI diagnostic tools used in the United States of America.

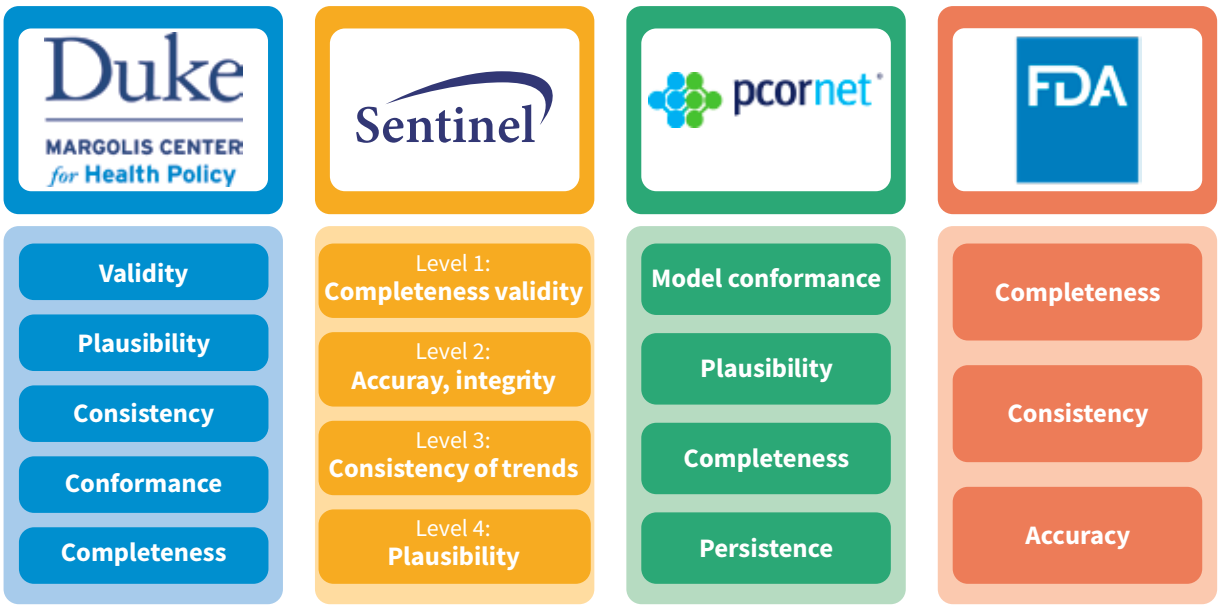


FIGURE 13. Examples of quality check principles *(41)*

- **Data labelling.** It is important to ensure consistent, reliable and accurate labelling of datasets for testing in line with good practices. In cases where subjective reference standards are used, quality will be influenced by many factors – such as the independence and qualifications of the graders, the number of graders per label, whether the reference standard is validated to correlate with patient outcomes, and whether the reference standard follows published metrics.

- **Documentation and transparency.** The algorithm and data supporting it are often not available or are not well documented for all AI system stakeholders. This makes it difficult to assess the quality of the underlying data. Transparency and careful documentation are important not only with regard to the methodology used in collecting data, but also for the selection and modifications of datasets used for training, validation and testing. Good documentation is fundamental to achieve transparency that enables verification and traceability. Transparency of methods should ensure data quality. Beyond the CONSORT-AI and SPIRIT-AI reporting guidelines, checklists have been devised by the machine learning community to report representativeness, completeness and other data quality characteristics *(44, 45).*

In addition, developers should consider deploying rigorous pre-release trials for AI systems to ensure that they will not amplify any of the issues discussed – such as biases and errors in the training data, algorithms, or other elements of system design. Furthermore, careful design or prompt troubleshooting can help identify data quality issues early. This could potentially prevent or mitigate possible resulting harm. Finally, to mitigate data quality issues that arise in health-care data and the associated risks, stakeholders should continue to work to create data ecosystems to facilitate the sharing of good-quality data sources.

The list in Table 3 summarizes the key data quality considerations for AI system safety and effectiveness.[5]

[5] This list will be updated and harmonized with the work of the IMDRF.

Table 3. General data quality considerations

Category	Data quality consideration item
Dataset	Splitting Selection volume and size Selection bias Individual variables definitions in each dataset Raw data versus “cleaned” data Data wrangling and cleansing Parameters and hyperparameters Usability Characterization Labelling Dependencies Augmentation Manipulation Streaming Interfaces Integrity Unique requirements Data source
Data infrastructure	Storage size Storage format Transformation medium
AI/ML model	Data training Tuning data Verification set Validation set Testing Development set Static AI versus dynamic AI Open AI versus closed AI
Governance management	Liability Data access Risk management Data protection Privacy Adoption education for clinical practice Good practices Standards (of care, governance, interoperability, etc.) Scope of practice and AI model use Technical checklist Documentation Transparency

PRIVACY AND DATA PROTECTION

5.5 Privacy and data protection

The WHO Global Strategy on Digital Health 2020–2025 classifies health data as sensitive personal data, or personally identifiable information, that requires a high standard of safety and security. Therefore, the strategy emphasizes the need for a strong legal and regulatory framework to protect the privacy, confidentiality, integrity, availability and processing of personal health data. A responsive legal and regulatory framework can also address issues of cybersecurity, trust-building, accountability and governance, ethics, equity, capacity-building and literacy. This will help ensure that good-quality data are collected and subsequently shared to support the planning, commissioning and transformation of services.

To develop and maintain adequate data security strategies, it is important for AI system developers, deployers and manufacturers to understand the thickening web of privacy and data protections laws. This section discusses high-level considerations for privacy and data protection. For other ethical considerations, refer to the deliverable of the Working Group on Ethical Considerations on AI for Health[6] *(46).*

5.5.1 Current landscape

As the demand for health-related data increases, the protection of privacy is creating a unique challenge for all stakeholders wishing to benefit from the many opportunities created by AI systems and technologies. One of the main reasons for this is that the high dimensionality of big data could make it difficult to apply anonymization and de-identification methods. Additionally, ensuring that large-scale datasets are secure from unauthorized access at each stage of the development process – collection, storage and management, transport, analysis, sharing and destruction – is an important consideration.

Some 145 countries and regions have data protection regulations and privacy laws that regulate the collection, use, disclosure and security of personal information *(47)*. There are many different definitions and interpretations of "data protection" and "privacy". In some cases, data protection and privacy are used interchangeably. However, although these concepts are similar and often overlap, their meanings are different, and developers should be aware of the legal and ethical implications that result from these differences.

Laws and regulations that cover "the management of personal information" are typically grouped under "privacy policy" in the United States and under "protection policy" in the European Union (EU) and elsewhere. These laws are often complex and may have conflicting obligations. When developing an AI system for therapeutic development or health-care applications, early in the development process the developers should consider gaining an understanding of applicable data protection regulations and privacy laws, including special regulatory provisions related to sensitive information such as genetic data. Developers should also consider national laws as well as regional ones. For instance, in the United States, although the Health Insurance Portability and Accountability Act (HIPAA) sets a baseline for protecting health data, states are empowered to enact stricter privacy laws (e.g. California's Consumer Privacy Act of 2018).

It is important to understand the jurisdictional scope of the various laws. For instance, because the scope of the GDPR is broad and its impact is significant, companies may want at least to evaluate the extent to which they are subject to it. Most privacy laws, including Singapore's Personal Data Protection Act, apply only to personal data processed within the country, whereas the GDPR[7] may apply to the personal data of EU citizens, regardless

[6] For a broader discussion of privacy and other ethical considerations for the use of AI, refer to the deliverable of the FG-AI4H's Working Group on Ethical Considerations on AI for Health and international, regional and national recommendations.

[7] See also India's proposed Personal Data Protection Act.

of the location where data are processed.[8] As a result, companies subject themselves to compliance obligations under the GDPR if they are located in the EU (including if any component of the organization is located in the EU), if they offer goods and services to individuals located in the EU, or if they monitor the behaviour of persons located in the EU.

It is also important for developers to understand the varied legal contexts and requirements for privacy-related concepts such as "identifiable," "anonymous" and "consent". For example, Chapter 1 of the United Kingdom of Great Britain and Northern Ireland's draft anonymization, pseudonymization and privacy-enhancing technologies guidance warns that referring to datasets as "anonymized" when they still may contain personal data in a pseudonymized form poses the risk of violating the United Kingdom of Great Britain and Northern Ireland's data protection law in the mistaken belief that the processing does not involve personal data *(48).* Consent requirements also vary according to the jurisdiction. For instance, various jurisdictions may require "explicit consent", with heightened information requirements for the processing of health-related data (GDPR Article 9) *(49)*. Therefore, developers may wish to consider the varied legal contexts when documenting how they address privacy-related concepts, including measures taken to meet consent requirements, and how they define anonymous or identifiable information.

In addition, certain jurisdictions have data protection regulatory frameworks that introduce reciprocity-based rules and place restrictions on the movement or transfer of data across borders. This may have a significant impact on the way in which data are processed and shared between countries. These provisions serve to curtail transnational data flows into and out of areas that are considered not to provide an "adequate" level of data protection.

Adequacy assessments may be required to determine whether a recipient country has thresholds of data protection laws and protections "essentially equivalent" or "substantially similar" to the jurisdiction from which the data were transferred. The GDPR, as a significant driver of emerging global data protection regimes, provides that the free transfer of personal data to third countries, non-European Union Member States, can primarily occur where the third country is considered by the EU Commission to have an "adequate" level of protection.[9] As of May 2023, the EU Commission had recognized only 13 countries as providing adequate protection *(50).*

Developers should be aware of the nuances of the different jurisdictions' regulations and laws and should consider documenting their data protection practices accordingly. In general, companies should consider keeping abreast of new laws and requirements, leveraging governance, risk analysis, policies, training and other strategies in a comprehensive and coherent way.

[8] Like the GDPR, the CCPA applies to natural persons who are California residents who are "domiciled in the state or who is outside the state for a temporary or transitory purpose". Cal. Code Regs. tit. 18, §17014.

[9] Data flows have increasingly become an important part of global interconnection and AI development. Although the Schrems II case pertains to the EU-US position on data transfers, the wider implications inform global data transfers and the way in which they are to be compatible with GDPR requirements, including the validity of standard contractual clauses which depend on whether effective mechanisms are in place to ensure compliance with the level of protection required under the GDPR. *Data Protection Commissioner v. Facebook Ireland Limited, Maximillian Schrems* (Case C-311/18, "Schrems II").

5.5.2 Documentation and transparency

Documentation and transparency are critical to facilitating trust with regard to privacy and data protection. Detailed privacy policy disclosures provide regulators with a benchmark by which to examine a company's handling of data. These disclosures should identify significant uses of personal information for algorithmic decisions. Depending on the jurisdiction, the disclosures may require the inclusion of other relevant information – e.g. the types and sources of health data collected and processed; the identities of the persons or organizations which determined the purpose or means of processing personal data; the identity of the person or organization which processed the data; the legal bases for processing the data; how the data were collected (including whether adequate notice was provided to the data subject and how consent requirements were met); and technical and organizational information on the storage of data, including security measures.

Developers must take privacy into account as they design and deploy AI systems. This includes designing, implementing and documenting approaches and methods to ensure a quality continuum across the development phases to protect data privacy *(49)*.[10] Privacy protections should not be limited only to addressing cybersecurity risks, especially since some privacy risks (e.g. harms to one's dignity which may cause embarrassment or stigma, or more tangible harms such as discrimination, economic loss or physical harm) (51) can also arise by means unrelated to cybersecurity incidents. Therefore, when developing solutions to address risks, developers should have a general understanding of the different origins of cybersecurity and privacy risks and should develop their risk management practices accordingly (Figure 14).

A compliance programme should consider risks and should develop privacy compliance priorities that take into account any specific potential harm as well as the enforcement environment. Developers may want to consider including in their documentation a description of the operations involved in the processing of personal data, a risk assessment, and the measures implemented to mitigate risks that take account of the interests of data subjects.

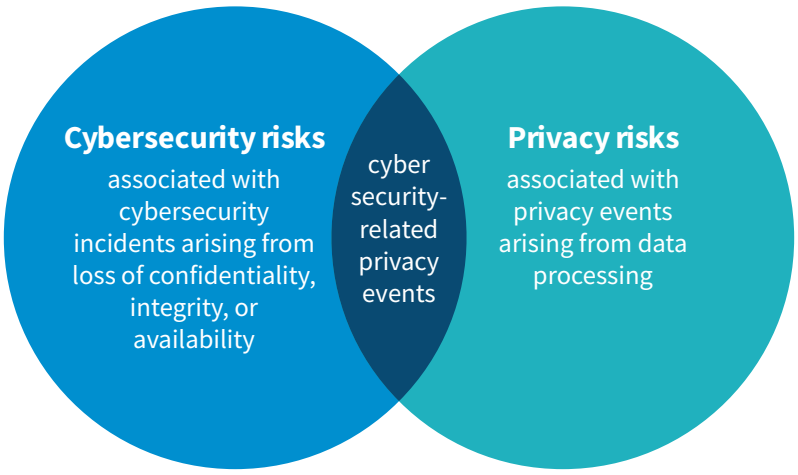


FIGURE 14. NIST Privacy Framework – cybersecurity and privacy risk relationship *(51)*

[10] For example, a pillar of the data quality continuum in some jurisdictions, e.g., EU law, is the accountability principle. According to Art. 5 of the GDPR, data controllers shall abide by the five sets of principles enshrined in Art. 5(1), e.g., data minimization. Data controllers shall determine both technical and organizational measures to attain such ends (Art. 5(2)), throughout the entire cycle of data processing. Although not mentioned, the accountability principle is also at work in Art. 24(1), 25(1), and 32 of the regulation in regard to the responsibility of the controller, principle of data protection by design (and by default), and security measures.

Certain regulations outline prescriptive security requirements to address cybersecurity and privacy risks – such as the GDPR's data protection by design and default (GDPR Articles 25 and 32) *(49)* and India's proposed data privacy by design policy (52) – while others include the duty to implement and maintain reasonable security practices and procedures appropriate to the risk.[11] Privacy frameworks often include privacy impact assessments, which are a widely used privacy management tool to proactively evaluate and mitigate privacy risks. Some jurisdictions, including the EU (GDPR Article 35) *(49)*[12], require companies to conduct these assessments.[13] Although United States of America's law does not require privacy impact assessments, the US Department of Commerce National Institute for Standards and Technology (NIST) privacy framework recommends that developers conduct them. According to NIST, "identifying if data processing could create problems for individuals, even when an organization may be fully compliant with applicable laws or regulations, can help with ethical decision-making in system, product, and service design or deployment" *(51).* This in turn can increase trust in the system.

Developers may also want to consider annotating their AI and having audit trails that explain what kinds of choices are made during the development process. Annotated notes provide "after the fact" transparency to outside parties and can help to explain the manner in which privacy was embedded, if applicable *(53).* Such explanations and documentation should be available at different levels of detail, targeted at different audiences – regulators, managers, developers, operators and users. The nature of the information and explanations required may differ, but all the assumptions, constraints, data sources, expected input and output, and major risks and limitations at each level should be clearly documented. In addition, an audit trail shows not only that controls have been applied but could also potentially show how damage was mitigated in the case of a data breach.

Many jurisdictions enforce certain cybersecurity requirements or publish guidance on cybersecurity for consideration by developers of medical devices. Although an in-depth discussion of cybersecurity requirements is outside the scope of this subsection, it is important to understand the key role that cybersecurity plays in the protection of personal health information. Cybersecurity focuses on specific technical implementations needed to protect systems and networks against cyberattacks, which could compromise both the security of health-related systems and data as well as an individual's privacy, which could result in harm. To provide transparency about cybersecurity practices, developers may wish to consider documenting practices and approaches for data security, including policies that help protect the confidentiality, integrity and availability of personal data throughout its lifecycle – such as appropriate encryption, access controls, logging methods, risk monitoring and methods of secure destruction. Developers may also consider documenting systems and approaches used to protect against data manipulation and adversarial attacks *(54)*. For instance, blockchain-based technologies may be one mechanism for protecting data privacy, security and integrity for AI in a traditionally fragmented health information systems ecosystem for national and regional contexts *(55).*

[11] For example: CCPA § 1798.150(a)(1), South Africa's Protection of Personal Information Act of 2013; Israeli Privacy Protection Regulations (Data Security), 5777–2017 (implementing the Protection of Privacy Law, 5741–1981 of 1981); United Arab Emirates' Federal Law No. 2 of 2019; Kingdom of Saudi Arabia's E-Commerce Law of 2019 and its Implementing Rules.

[12] "A data protection impact assessment shall be conducted if processing is likely to result in high risk to the rights and freedoms of the natural persons".

[13] While risk assessments are quite common in information security standards and requirements, they are rarely seen in privacy rules in the United States of America. The GDPR, however, requires that an organization processing personal data must conduct a specific Data Privacy Impact Assessment or DPIA before beginning the processing.

5.5.3 AI regulatory sandboxes

The above regulatory challenges are recognized by regulatory authorities and policy-makers across the world *(56)*. As a result, over 50 countries are currently experimenting with sandboxes in a wide range of high-technology sectors – notably in the financial sector but sandboxes have also gained popularity for health and legal services *(57)*. The regulatory sandbox approach has gained considerable traction as a means of helping regulators to address the development and use of AI and other emerging technologies *(57)*. Regulatory sandboxes are generally regulatory tools that allow the flexibility to test innovative products or services with minimal regulatory requirements *(57)*. Consequently, regulatory sandboxes are considered an agile approach to innovation and regulation and thus regulatory authorities are increasingly favouring them. In the EU, regulatory sandboxes have been proposed for testing surveillance solutions in the fight against the COVID-19 pandemic, and for establishing a framework for EU-wide data access. In relation to AI regulations specifically, the first AI regulatory sandbox pilot presumably launched in 2023 by the Government of Spain with an aim to provide a guide to all EU Member States and the European Commission *(58)*. Although AI regulatory sandboxes raised a few concerns, they have the potential to bring many key benefits to AI system regulators, developers, manufacturers and even patients *(57).* This is because such AI regulatory sandboxes can: 1) help enable a better understanding of the AI systems during the development phase and before they are placed on the market; 2) facilitate the development of adequate enforcement policies and technical guidance that can mitigate risks and unintended consequences; and 3) foster AI innovation by establishing a controlled experimentation and testing environment for innovative AI technologies, products and services for new and potentially safer AI systems.

ENGAGEMENT AND COLLABORATION

5.6 Engagement and collaboration

Where applicable and appropriate, engagement and collaboration between developers, manufacturers, health-care practitioners, patients, patient advocates, policy-makers, regulatory bodies and other stakeholders can improve the safety and quality of an AI system. Many regulatory bodies have adopted engagement and collaborative approaches in this area, and this section discusses the approaches of five of them: the United Kingdom of Great Britain and Northern Ireland's MHRA, the South African Health Products Regulatory Authority (SAHPRA), the European Commission, Singapore's HSA, and the U.S. FDA. Table 4 lists examples of with whom, why and how these regulators foster engagement and collaboration. The examples are not meant to be comprehensive but instead are intended to highlight general approaches. Table 4 is followed by an analysis that discusses the similarities and differences in the approaches.

Subsection 5.6.2 examines two examples of engagement and communication between regulators and AI developers resulting in positive clinical outcomes (CURATE.AI and IDentif.AI). The last subsections consider the practical implications for engagement and collaboration in resource-limited settings and recommend ways that regulatory bodies can initiate this process even in countries without past experience in engagement and collaboration. This is supplemented by several narratives: how to apply engagement tools (based on experience) and how to position the regulator as a partner in the context of accessible dialogue, and guidance and recommendations during the development process.

Table 4. Examples of regulators' approaches to engagement and collaboration with stakeholders about the use of AI in health care and therapeutic development

1. Medicines and Healthcare Products Regulatory Agency (MHRA), United Kingdom of Great Britain and Northern Ireland	
With whom?	**Examples of stakeholders with whom the MHRA engages and collaborates:** • **Patients/patient advocates** • **Academia** • **Health-care professionals** e.g. providers in the National Health Service (NHS) and private health-care providers. • **Industry** e.g. medical device and in vitro diagnostics industry, health technology industry. • **Domestic government partners** e.g. Department of Health and Social Care (DHSC), NHS England and Improvement, NICE, and Care Quality Commission (CQC).
Why?	**Examples of reasons why the MHRA engages and collaborates with stakeholders:** • Alert users to problems with medical devices and medicines. • Answer enquiries about roles in regulation or raise awareness of safety issues. • Seek feedback on development of regulatory policy, managing adverse incidents and risks. • Interface with United Kingdom of Great Britain and Northern Ireland government and NHS, including stakeholders aligned to digital and AI-related activities.
How?	**Examples of ways in which the MHRA engages and collaborates with stakeholders:** • Central alerting system to the NHS and health-care providers or through professional groups. • Media, public, and other stakeholder inquiries via MHRA customer service centre, dedicated email inboxes, and press office. • Connecting with expert advisory groups, networks, and stakeholder groups on specific issues. • Consultation on engagement with patients and public *(59)*. • Working-level meetings with national stakeholders, bilateral meetings with other parts of NHS, government and international counterparts.

Table 4. Examples of regulators' approaches to engagement and collaboration with stakeholders about the use of AI in health care and therapeutic development, cont.

2. South African Health Products Regulatory Authority (SAHPRA), South Africa	
With whom?	**Examples of stakeholders with whom SAHPRA engages and collaborates:** • **Patients/patient advocates** • **Academia** • **Health-care professionals** • **Industry** (e.g. manufacturers/ distributors, trade associations). • **National government partners** (e.g. National Department of Health, National Department of Trade & Industry, South African National Accreditation Service).
Why?	**Examples of reasons why the SAHPRA engages and collaborates with stakeholders:** • Facilitate the approval of innovative AI systems. • South African National Accreditation System (SANAS) to ensure that the Conformity Assessment Body network is established in the country to certify the quality management system (QMS)
How?	**Examples of ways in which the SAHPRA engages and collaborates with stakeholders:** • The framework for engagement and collaboration has not yet been formalized. • Recommended that stakeholder engagement adopt the five-step engagement model developed by TGA (60).
3. EC (European Union)	
With whom?	**Examples of stakeholders with whom the EC engages and collaborates:** • **Patients/patient advocates** • **Academia** • **Health-care professionals**
Why?	**Examples of reasons why the EC engages and collaborates with stakeholders:** • To "support the Commission in the development of actions for the digital transformation of health and care in the EU."
How?	**Examples of ways in which the EC engages and collaborates with stakeholders:** • By providing "advice and expertise to the Commission, particularly on topics set out in the communication (61) on enabling the digital transformation of health and care in the Digital Single Market, that was adopted in April 2018." In particular, such topics regard health data interoperability and record exchange formats, digital health services, data protection and privacy, AI, and "other cross cutting elements linked to the digital transformation of health and care, such as financing and investment proposals and enabling technologies."

Table 4. Examples of regulators' approaches to engagement and collaboration with stakeholders about the use of AI in health care and therapeutic development, cont.

4. Health Sciences Authority (HSA), Singapore	
With whom?	**Examples of stakeholders with whom the HSA engages and collaborates:** • **Academia** (e.g. research institutions). • **Health-care professionals** • **Industry** (e.g. software and AI developers, trade associations). • **National government bodies**
Why?	**Examples of reasons why the HSA engages and collaborates with stakeholders:** • Early engagement and support to innovators to facilitate regulatory compliance, thus facilitating timely access to safe innovations for patients. • Actively consult on new policies and guidelines related to AI and software medical devices to receive and incorporate stakeholders' inputs and perspectives (Regulatory guidelines for software medical devices – a life cycle approach *(16)*. • To work with other agencies responsible for implementation and deployment of AI and software medical devices in the health-care system to facilitate greater adoption of innovative technologies in the health-care system.
How?	• Rapid, streamlined engagement portals are available for several facets of product regulation *(62)*. • Specific processes that can be straightforwardly addressed include Medical Device Information Communication System (for application submissions for licences, permits, registrations, etc.). • Online self-help tools to determine the product classification and risk classification for medical devices and simple forms to seek advice and confirmation from the HSA. • **Medical Device Development Consultation:** Online appointment booking system that allows innovators and developers to seek scientific and regulatory advice during the medical device development phase to facilitate regulatory compliance. • Online stakeholder consultation process for all new and revised policies and guidelines. • Regular focus group discussions and engagements with industry associations and companies.

Table 4. Examples of regulators' approaches to engagement and collaboration with stakeholders about the use of AI in health care and therapeutic development, cont.

5. Food and Drug Administration (FDA), United States of America	
With whom?	**Examples of stakeholders with whom the FDA engages and collaborates:** • **Patients/caregivers/patient advocates** • **Academia** (e.g. research institutions). • **Health-care professionals** • **Industry** (e.g. developers, device manufacturers, drug companies, trade associations). • **National government partners** (e.g. National Institutes of Health [NIH], Office of the National Coordinator for Health Information Technology [ONC], Federal Communications Commission [FCC]). • **Foreign government partners** • **International organizations** (e.g. IMDRF, ICH). • **Consumers/general public**
Why?	**Examples of reasons why the FDA engages and collaborates with stakeholders:** • Facilitate patient access to technologies that can benefit them in a timely manner. • Support novel, innovative medical product development through early interactions with stakeholders. • Provide timely feedback on FDA policies to reduce uncertainty. • Communicate to the public about AI/ML devices. • Receive feedback on policies, guidance and discussion papers.
How?	**Examples of ways in which the FDA engages and collaborates with stakeholders:** • Hold different types of pre-submission meetings to provide early feedback to sponsors. • Participate and lead international harmonization efforts (e.g. IMDRF, ICH). • Engage as members of public-private partnerships and collaborative communities. • Collaborate in pre-competitive space on regulatory science research to advance scientific community understanding. • Receive formal comments on policies and guidance through the Federal Register. • Hold workshops and other engagement events to obtain feedback from patients, industry and other stakeholders.

5.6.1 Discussion on strategies of profiled regulatory bodies

Table 4 shows the approaches of four national and one regional (in the case of the EC) regulatory body to foster engagement and collaboration. In the first category ("with whom?"), there are considerable similarities between these bodies. The shared targets for engagement and collaboration include health professionals (indicated by FDA, SAHPRA, MHRA, EC and HSA), academia (FDA, SAHPRA, MHRA, EC and HSA), industry (FDA, SAHPRA, MHRA, EC and HSA), patients or patient advocates (FDA, SAHPRA, MHRA and EC), domestic government bodies (FDA, SAHPRA and MHRA), media (national and trade press; FDA and MHRA), health providers (FDA and MHRA) and consumers (FDA and MHRA). Interestingly, the strategy paper by the US Department of Commerce's NIST also refers to academia and domestic government bodies as targets for engagement and collaboration.

In the second category ("why?"), SAHPRA notes the importance of communicating the benefits and intended use of devices, presumably to protect and promote public health (listed by the FDA and implied by MHRA). The FDA also stresses the importance of bilateral communication with stakeholders so that regulators are aware of developments in industry (or academia) and so that these stakeholders, in turn, are aware of developments in regulation. Similarly, MHRA indicates the importance of acquiring feedback about medical devices from

stakeholders. This supports the objectives given by both SAHPRA and the EC, namely to facilitate approval of innovative solutions and support the digital transformation of health and care. The HSA acknowledges the importance of early engagement with innovators and developers to provide greater clarity in regulatory requirements and improve transparency in regulatory processes.

For the third category ("how?"), the FDA lists steps that are taken to foster engagement (e.g. hosting workshops, producing digital and print material, and offering training modules or other types of education). MHRA also notes the importance of holding meetings with stakeholders (including domestic government institutes and international counterparts). HSA has introduced a pre-market consultation scheme to support innovation and device development by providing scientific and regulatory advice to enable regulatory compliance by software and AI developers who, unlike traditional medical device manufacturers, are not familiar with regulatory requirements *(60, 63).*

5.6.2 Two successful instances of engagement

To understand the value of engagement and collaboration between regulatory bodies and stakeholders, two real-world examples (Case 1 and Case 2) are described. Clear avenues for engagement between regulators and AI developers play a major role in ensuring that rigorous evaluation and accelerated delivery of impactful modalities can be realized seamlessly. One aspect is in the area of interventional AI/digital medicine, which involves the application of software/devices (e.g. AI-based drug development and/or dosing platforms) and/or the application of resulting drug compounds and/or combinations recommended by these platforms *(64, 65, 66)*. In this context, integrating regulator accessibility with emerging innovation, sometimes in urgent circumstances, will ultimately result in life-saving outcomes. Importantly, these outcomes will not be confined to post-approval treatment but also to substantial patient benefit during the investigational stages of validation.

In **Case 1**, the developmental roadmap and validation of CURATE.AI and foundational technology of IDentif.AI were discussed with the Medical Devices Branch *(16)* of the HSA in Singapore. This interactive session included an in-depth review of the key findings of the technology platforms, the process of implementing both platforms, emerging statistical analysis strategies to assess effectively the personalized medicine treatment outcomes and regulatory routes. A broader discussion on how clinical trial designs may evolve due to the emergence of AI was also conducted *(68, 69, 70)*. A clear pathway for subsequent inquiries was established, as multiple and frequent guidance requests were expected due to the nature of the trial designs that were envisioned. These included N-of-1 study designs for a broad range of indications designed for each patient. Specifically, these designs were personalized on the basis of (for example) the individualized dosage calibrations of the drug regimen (clinician-selected regimen), serial efficacy and toxicity measurements, efficacy-guided treatment protocols, and safety parameters. Subsequent submissions have included engagement with regulators for risk classifications associated with the device for each trial and subsequent discussion for submission of Special Access Routes (SARs) *(71)* for the potential rapid implementation of trials and for treatment purposes if needed. Rapid and informative responses and active engagement from HSA regulatory team members resulted in efficient turnaround times for trial initiation, which ultimately resulted in a positive outcome for a refractory oncology patient. A sustained track record of engagement with the regulatory community has played a key role in helping a clear process flow to be developed for downstream guidance requests.

Case 2 was developed in response to the COVID-19 pandemic. Specifically, a patient-derived live virus strain was harnessed for IDentif.AI-driven combination therapy optimization to serve as a clinical decision support system (CDSS). Unlike traditional AI-based approaches, this strategy did not use existing patient datasets. Instead, prospective experimentation was used alongside an AI-derived small data analytics strategy to pinpoint prospective data-backed rankings of combinations for potential further clinical consideration and

potentially for the elimination of certain combinations from further clinical consideration. The foundational technology for IDentif.AI was previously discussed in detail with the HSA Medical Devices Branch, and additional IDentif.AI SARS-CoV-2 study information was provided in the context of clinical decision support, developing optimized combinations pinpointed by IDentif.AI and with potential trials being designed with clinical partners. With regard to regulator engagement, the Medical Devices Branch of the HSA was contacted to provide device risk classification guidance for the submission of a Clinical Research Materials Notification (CRM-N) for study purposes. Obtaining a CRM-N is a required part of the submission of a clinical validation programme because it stipulates the prerequisite of an initial assessment of device risk from the HSA *(72)*. The submission portal and portal interaction were particularly straightforward to navigate and were integrated with a uniform access portal which was streamlined for efficient oversight and monitoring with regulatory bodies. This further demonstrates the straightforward process of interaction with the HSA. This case was an example of the critical importance of straightforward regulator accessibility and the profoundly positive impact that this can have on the advancement of promising technologies towards further clinical assessment and validation.

5.6.3 Recommended approaches for countries without past experience

For countries with limited experience in engagement and collaboration (and/or limited resources), it is important to establish: 1) what levels of engagement and collaboration are desired; 2) what steps can and should be taken to achieve those levels; and 3) what challenges are presented by the technology (e.g. AI explainability).

In many cases, it is desirable to adopt regulatory models that are adaptable, flexible, modular and scalable in order to account for the uncertainties of innovation through appropriate oversight and coordination. These features fit not only the specific challenges of emerging technologies but also of the regulatory approach of countries without past experience in this field or with scarce economic resources. On the one hand, priorities should be scalable so that growing amounts of work can be suitably addressed by adding resources to the regulatory model. On the other hand, however, priorities should be determined in accordance with the modular adaptability of the steps and levels of engagement. In ecology, adaptability applies to the ability to cope with unexpected disturbances in the environment. In engineering, modularity refers to the interrelation of the separate parts of a software package or to the partitioning of the design to make it manageable. In multi-agent systems (MAS), it refers to the efficient usage of computational resources. We can profit from this notion to create adaptable policies that can be combined into regulatory systems for legal governance. The aim should be to address the uncertainties of innovation and to align with society's preferences on emerging innovation, while allowing regulators to gain a growing understanding of technological challenges with increasing normative granularity *(73).*

5.6.4 Narrative on using engagement tools based on practical experience

For all countries – from those with limited experience in engagement and collaboration (and/or limited resources) to those at the other end of the spectrum – project and programme management tools can help organizations (including regulators) to structure and execute their engagement with stakeholders and users. No matter which tool is chosen, the key to valuable engagement is to invest time, energy and thought into how best to engage stakeholders and then following through on that engagement for the duration of a project or programme. Engagement often fails if the investment is seen as a short-term rather than long-term relationship.

The Australian Government's recommended five-step model for engagement *(60)* is a good starting point for considering how a regulator could engage with developers of AI health products and services. In this model, engagement starts with thinking through the purpose of the engagement (based on what it is hoped to achieve) and identifying the relevant stakeholders. When planning the different levels of engagement with stakeholders,

it is recommended to map out existing relationships and to define the type of engagement and relationship that is needed with the stakeholder (and what type of relationship the stakeholder would be open to having). For instance, a digital health developer building an application (app) to support parents with children above a healthy weight may find that the primary health body concerned is an influential stakeholder which sets policies on managing children's weight. However, this is not a body with whom the developer of the app needs to engage regularly, so the developer may only "inform" the health body of the project. However, a developer will want to work with parents of children above a healthy weight to co-design the app and ensure that it fits their needs. It would, therefore, be important for the developer to "collaborate" with a representative group of parents and establish two-way or multi-way communication and shared learning and decision-making over the course of the project.

A similar approach for making sure that stakeholders are provided with the right information at the right time and are using optimal communication channels is outlined by one of the leading product development software companies *(74).* Within the stakeholder communication "play", importance is placed on who the stakeholders are, the desired method of communication and the frequency of communication. For instance, an internal government project developing a digital health product will have internal stakeholders (such as funders of the project and policy leads) and external stakeholders (such as leading academics). The communications plan should outline how each stakeholder group will be addressed (email, face-to-face conversation, video call, and/or social media) and how often there were will be contact with the stakeholder group (daily, fortnightly, and/or yearly) based on what the relationship with the stakeholder brings to the overall goals (i.e. information-sharing, co-design, and/or quality assurance). This plan can then be mapped out in a simple table (for which examples of headings might be: method, audience/stakeholder, content to share, why, and frequency) for the whole development team to follow.

5.6.5 Narrative positioning the regulator as a partner in the development process

As demonstrated in Table 4 and discussed in the subsequent text, multiple regulatory bodies emphasize the importance of open (bilateral) communication with stakeholders so that regulators are aware of developments in AI-based technology and so that these stakeholders, in turn, are aware of changes in regulation. This is because AI-based technology is constantly changing and regulation needs to be able to keep pace. The development, deployment, post-market surveillance and iteration of AI products and services in health care should therefore be an ongoing conversation between developers and regulators.

It is recommended that regulators look at AI-based technology in health care from a mindset of accessible engagement that potentially, when applicable, facilitates working alongside the developer to ensure compliance with regulatory requirements throughout the development and implementation process. An engagement mindset approach to regulation is about building trusting, collaborative relationships between developers and the regulatory body(s), and a two-way dialogue that enables developers to learn from regulators and vice/versa.

Furthermore, depending on a country's regulatory arrangements, one or more regulators may be responsible for AI-based health products and services. This means a developer often has to work with (and meet the standards of) more than one regulatory body. To ensure that this is a smooth and positive experience for AI developers, it is again recommended that regulators take a service approach. This means that a single, clearly marked pathway should be established and should be followed by an AI developer when ensuring the compliance of a product or service. Regulators need to collaborate with each other on issues such as clear messaging to developers and consistent levels of engagement with developers at the right point, and by sharing what they learn from different engagements with developers.

If a country wishes to take an accessible engagement approach to the regulation of AI products and services, co-regulation could be explored. As outlined by Clarke *(75)*, in a co-regulation approach regulators outlined a regulatory framework based on required compliance to the legislative act(s). The details of how this is applied in practice are jointly developed by regulators and a representative sample of developers *(75)*. Similarly, when considering regulation from a service mindset, a co-regulatory approach, when appropriate and with any potential conflicts of interest properly managed, is about generating buy-in from developers by engaging them in the design and implementation of the regulatory process. The approach involves designing a regulatory process that reflects and acknowledges the needs of developers and not just those of the regulatory body and associated groups. Ultimately, however, regulators must remain fully independent of developers in order to make decisions that put the safety of the public first, as well as ensuring that public and private health-care resources are used only for technologies that meet independently developed standards of quality, safety and efficacy.

6. RECOMMENDATIONS FOR THE WAY FORWARD

Based on its work, the WG-RC recommends that stakeholders examine the key 18 considerations discussed in Section 5 above and summarized in Table 5 below as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development.

TABLE 5. Key recommendations for regulatory considerations on AI for health based on each of the six topic areas

1. Documentation and transparency recommendations
1.1 Consider pre-specifying and documenting the intended medical purpose and development process, such as the selection and use of datasets, reference standards, parameters, metrics, deviations from original plans, and updates/changes during the phases of development. These should be considered in a manner that allows for the tracing of the development steps, as appropriate.
1.2 Consider a risk-based approach also for the level of documentation and record-keeping utilized for the development and validation of AI systems.
2. Risk management and AI systems development lifecycle approach recommendations
2.1 Consider a total product lifecycle approach throughout all phases in the life of a medical device: pre-market development management, post-market management/surveillance, and change management.
2.2 Consider a risk management approach that addresses risks associated with AI systems, such as cybersecurity threats and vulnerabilities, underfitting, algorithmic bias etc.
3. Intended use, and analytical and clinical validation recommendations
3.1 Consider providing transparent documentation of the intended use of the AI system. Details of the training dataset composition underpinning an AI system – including size, setting and population, input and output data and demographic composition – should be transparently documented and provided to users.
3.2 Consider demonstrating performance beyond the training dataset through external, analytical validation in an independent dataset. This external validation dataset should be representative of the population and setting in which the AI system is intended to be deployed and transparent documentation of the external validation dataset and performance metrics should be provided. This external validation dataset should be appropriately independent of the dataset used for the development of the AI model during training and testing.
3.3 Consider a graded set of requirements for clinical validation based on risk. Randomized clinical trials are the gold standard for the evaluation of comparative clinical performance and could be appropriate for the highest risk tools or where the highest standard of evidence is required. In other situations, consider prospective validation in a real-world deployment and implementation trial which includes a relevant comparator using accepted relevant groups.
3.4 Consider a period of more intense post-deployment monitoring through post-market management and market surveillance for high-risk AI systems.

TABLE 5. Key recommendations for regulatory considerations on AI for health based on each of the six topic areas, cont.

4. Data quality recommendations
4.1 Consider whether available data are of sufficient quality to support the development of the AI system that can achieve the intended purpose.
4.2 Consider deploying rigorous pre-release evaluations for AI systems to ensure that they will not amplify any of relevant issues, such as biases and errors.
4.3 Consider careful design or prompt troubleshooting to help early identification of data quality issues, which could potentially prevent or mitigate possible resulting harm.
4.4 Consider mitigating data quality issues that arise in health-care data and the associated risks.
4.5 Consider working with other stakeholders to create data ecosystems that can facilitate the sharing of good-quality data sources.
5. Privacy and data protection recommendations
5.1 Consider privacy and data protection during the design and deployment of AI systems.
5.2 Consider gaining a good understanding of applicable data protection regulations and privacy laws early in the development process and ensure that the development process meets or exceeds such legal requirements.
5.3 Consider implementing a compliance programme that addresses risks and develop privacy and cybersecurity practices and priorities that take into account potential harm and the enforcement environment.
6. Engagement and collaboration recommendations
6.1 Consider the development of accessible and informative platforms that facilitate engagement and collaboration, where applicable and appropriate, among key stakeholders of the AI innovation and deployment roadmap. and collaboration
6.2 Consider streamlining the oversight process for AI regulation through engagement and collaboration in order potentially to accelerate practice-changing advances in AI.

7. CONCLUSION

WHO recognizes the potential of AI in enhancing health outcomes by improving clinical trials, medical diagnosis, treatment, self-management of care and person-centred care, as well as creating more evidence-based knowledge, skills and competence for professionals to support health care. Furthermore, with the increasing availability of health-care data and the rapid progress of analytics techniques, AI has the potential to transform the health sector to meet a variety of stakeholders' needs in health care and therapeutic development. For this reason, WHO and ITU are collaborating through the Focus Group on AI for Health (FG-AI4H) to facilitate the safe and appropriate development and use of AI systems in health care. The FG-AI4H's Working Group on Regulatory Considerations (WG-RC) on AI for Health consists of members representing multiple stakeholders – including regulatory bodies, policy-makers, academia and industry – who explored regulatory and health technology assessment considerations and emerging "good practices" for the development and use of AI in health care and therapeutic development. This publication, which is based on the work of the WG-RC, is an overview of regulatory considerations on AI for health that covers the following six general topic areas: Documentation and transparency, Risk management and the AI Systems Development Lifecycle Approach, Intended use and analytical and clinical validation, Data quality, Privacy and data protection, and Engagement and collaboration. This overview is not intended as guidance, regulation or policy. Rather, it is a list of key regulatory considerations and is a resource that can be considered by all relevant stakeholders in medical devices ecosystems, including developers who are exploring and developing AI systems, regulators who might be in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI-embedded medical devices, health practitioners who deploy and use such medical devices and AI systems, and those working in this area. The WG-RC recommends that stakeholders examine these key considerations and other potential ones as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development in relationship to the 6 topic areas.

The WG-RC recognizes that AI has been instrumental in rapidly advancing research in health care and therapeutic development. However, it also recognizes the evolving complexity of the AI landscape and the need for international collaboration to facilitate the safe and appropriate development and use of AI systems. Accordingly, international collaboration on AI regulations and standards is important for three reasons. First, sharing knowledge and best practices of evolving regulatory considerations could increase the speed of developing this regulatory landscape and reduce the gap between advancing technology and regulation. Second, international collaboration improves consistency in regulations, which is important as many tools are likely eventually to cross borders. Consistency of regulatory considerations for AI systems and technologies could improve standards and enable more rapid deployment. Third, international collaboration supports countries with less regulatory capacity by ensuring that these countries can also use tools with high standards, reducing the potential for disparity in the introduction of these tools. Eventually, the WG-RC understands that the AI landscape is rapidly evolving and that the considerations in this deliverable may need to be expanded as the technology and its uses develop. The working group recommends that stakeholders, including regulators and developers and manufacturers, continue to engage and that the community at large works towards shared understanding and mutual learning. In addition, established national and international groups, such as the IMDRF, GHWP, AMDF and ICMRA, should continue to work on AI topics for potential regulatory convergence and harmonization.

REFERENCES

1. Global Strategy on Digital Health 2020–2025. Geneva: World Health Organization; 2020 (https://apps.who.int/iris/handle/10665/344249, accessed 25 July 2023).

2. The 17 Goals – Sustainable Development (online). New York (NY): United Nations; 2020. (https://sdgs.un.org/goals, accessed 25 July 2023).

3. Thirteenth General Programme of Work 2019−2023. Geneva: World Health Organization (https://www.who.int/about/what-we-do/thirteenth-general-programme-of-work-2019---2023, accessed 25 July 2023).

4. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Discussion paper and request for feedback. Silver Spring (MD): US Food and Drug Administration; 2019 (https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf, accessed 25 July 2023).

5. Informal innovation network. Horizon scanning assessment report – Artificial Intelligence. International Coalition of Medicines Regulatory Authorities; 2021 (https://www.icmra.info/drupal/sites/default/files/2021-08/horizon_scanning_report_artificial_intelligence.pdf, accessed 25 July 2023).

6. ISO/IEC TR 24028:2020, Information technology – artificial intelligence – overview of trustworthiness in artificial intelligence (https://www.iso.org/standard/77608.html, accessed 25 July 2023).

7. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Paris: Organisation for Economic Co-operation and Development; 2019 (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, accessed 25 July 2023

8. Machine learning-enabled medical devices: a subset of AI-enabled medical devices: key terms and definitions. Proposed document posted for public consultation, 16 September 2021. International Medical Device Regulators Forum; 2021 (https://www.imdrf.org/sites/default/files/2021-10/Machine%20Learning-enabled%20Medical%20Devices%20-%20A%20subset%20of%20Artificial%20Intelligence-enabled%20Medical%20Devices%20-%20Key%20Terms%20and%20Definitions.pdf, accessed 25 July 2023).

9. Panesar A. Machine learning and AI for healthcare. Big data for improved health outcomes. Coventry: Apress; 2019.

10. Artificial intelligence and intellectual property policy (online). Geneva: World Intellectual Property Organization; 2022 (https://www.wipo.int/about-ip/en/artificial_intelligence/policy.html, accessed 3 July 2023).

11. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582–4.

12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74.

13. Rivera SC, Liu X, Chan A, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ 2020;370:m3210.

14. Guidance for post-market surveillance and market surveillance of medical devices, including in vitro diagnostics. Geneva: World Health Organization; 2020 (https://apps.who.int/iris/handle/10665/337551, accessed 25 July 2023).

15. Software as a Medical Device (SaMD): key definitions. International Medical Device Regulators Forum; 2013. (http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf, accessed 25 July 2023).

16. Regulatory guidelines for software medical devices – a lifecycle approach (online). Singapore: Health Sciences Authority; 2022 (https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach_r2-(2022-apr)-pub.pdf, accessed 25 July 2023

17. Oala L, Heiß C, Macdonald J, März M, Kutyniok G, Samek W. Detecting failure modes in image reconstructions with interval neural network uncertainty. Int J Comput Assist Radiol Surg. 2021;16(12):2089–97.

18. Oala L, Johner C, Goldschmidt P.G., Balachandran P. Good Practices for Health Applications of Machine Learning: Considerations for Manufacturers and Regulators. In: Proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) – Meeting O; 2023 (https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-AI4H-2022-2-PDF-E.pdf , accessed 25 July 2023).

19. Principles and practices for medical device cybersecurity. International Medical Device Regulators Forum; 2019 (http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-ppmdc.pdf, accessed 25 July 2023).

20. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Action plan. US Food and Drug Administration; 2021 (https://www.fda.gov/media/145022/download, accessed 25 July 2023).

21. Software as a medical device: possible framework for risk categorization and corresponding considerations. International Medical Device Regulators Forum; 2014 (https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf, accessed 25 July 2023).

22. A buyer's guide to AI in health and care. London: NHSX; 2020 (https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/, accessed 25 July 2023).

23. Notification No.0831-14, 31 August 2020 (Chinese). Handling with applications for confirmation of PACMP for medical devices, PSEHB/SD (in Japanese). Tokyo: Ministry of Health, Labour and Welfare; 2020 (https://www.mhlw.go.jp/content/11120000/000665757.pdf, accessed 25 July 2023).

24. Workshop on clinical evaluation of AI for health. Geneva: International Telecommunication Union; 2020 (https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/ws/2010.aspx, accessed 25 July 2023).

25. Schörverth E, et.al. FG-AI4H Open Code Initiative – evaluation and reporting package. In: proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) – Meeting K; 2021.

26. Sendak M-P, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ digital medicine. 2020;3(1):1–4.

27. Verks B, Oala L. Data and artificial intelligence assessment methods (DAISAM) Audit Reporting Template. In: Proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) – Meeting J, 2020.

28. Oala L, Fehr J, Gilli L, Balachandran P, Leite AW, Calderon-Ramirez S et al. ML4H Auditing: from paper to practice. In: Proceedings of Machine Learning for Health (ML4H) NeurIPS Workshop. Proceedings of Machine Learning Research. 136:280-317. (https://proceedings.mlr.press/v136/oala20a.html, accessed 25 July 2023).

29. Willis K, Oala L. Post-hoc domain adaptation via guided data homogenization. (https://arxiv.org/abs/2104.03624, accessed 25 July 2023).

30. Calderon-Ramirez S, Oala L. More than meets the eye: semi-supervised learning under non-IID data. Presented as a RobustML workshop paper at International Conference on Learning Representations (ICLR), 2021 (https://arxiv.org/abs/2104.10223, accessed 25 July 2023).

31. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health. 2019;1(1):e35–e44.

32. Macdonald J, März M, Oala L, Samek W. Interval neural networks as instability detectors for image reconstructions. In: Palm C, Deserno TM, Handels H, Maier A, Maier-Hein K, Tolxdorff T, editors. Bildverarbeitung für die Medizin. Informatik aktuell (Image processing for medicine. IT update). Wiesbaden: Springer Vieweg; 2021.

33. International Digital Health and AI Research Collaborative (I-DAIR) (online) (http://i-dair.org/, accessed 25 July 2023).

34. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. 2020;6(10):1581–8.

35. FG-AI4H Open Code Initiative (OCI). International Telecommunication Union; 2022 (https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/opencode.aspx, accessed 16 March 2023).

36. AI audit.org (website) (https://aiaudit.org/, accessed 25 July 2023).

37. Software as a medical device (SaMD): clinical evaluation. International Medical Device Regulators Forum; 2016 (http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-samd-ce.pdf, accessed 25 July 2023).

38. Evidence standards framework for digital health technologies. London: National Institute for Health and Care Excellence (NICE); 2019 (https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf, accessed 25 July 2023).

39. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med. 2020;26(9):1318–20.

40. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Draft guidance for industry. Silver Spring (MD): US Food and Drug Administration; 2021 (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory, accessed 25 July 2023).

41. Determining real-world data's fitness for use and the role of reliability. Durham (NC): Duke-Margolis Center for Health Policy; 2019 (https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf, accessed 25 July 2023

42. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.

43. Shana L. The geographic bias in medical AI tools. Ethics and Justice, Healthcare, Machine Learning. Stanford (CA): Stanford University Human-Centered Artificial Intelligence News and Announcements, 21 September 2020 (https://hai.stanford.edu/news/geographic-bias-medical-ai-tools, accessed 25 July 2023).

44. The dataset nutrition label (online). The Data Nutrition Project (https://datanutrition.org/, accessed 25 July 2023).

45. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86–92.

46. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. (https://apps.who.int/iris/handle/10665/341996, accessed 25 July 2023).

47. Greenleaf G. Global tables of data privacy laws and bills, seventh edition (February 11, 2021) 169 Privacy Laws & Business International Report; 2021:6–19. (https://ssrn.com/abstract=3836261 or http://dx.doi.org/10.2139/ssrn.3836261, accessed 25 July 2023).

48. Introduction to anonymisation: draft anonymisation, pseudonymisation, and privacy enhancing technologies guidance. London: Information Commissioner's Office (ICO); 2021 (https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf, accessed 25 July 2023).

49. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

50. Adequacy decisions: how the EU determines if a non-EU country has an adequate level of data protection. Brussels: European Commission (https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en, accessed 25 July 2023).

51. NIST privacy framework: a tool for improving privacy through enterprise risk management. Washington (DC): National Institute of Standards and Technology (NIST), US Department of Commerce; 2020 (https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf, accessed 25 July 2023).

52. India's Personal Data Protection Act. Chapter VI, 22(1)(e), 24(1).

53. West DM, Allen JR. Turning point: policymaking in the era of artificial intelligence. Washington (DC): Brookings Institution Press; 2020.

54. Framework for improving critical infrastructure cybersecurity. Washington (DC): National Institute of Standards and Technology (NIST), U.S. Department of Commerce; 2018 (https://www.nist.gov/cyberframework, accessed 25 July 2023).

55. Alsalamah SA, Alsalamah HA, Nouh T, Alsalamah SA. *HealthyBlockchain* for global patients. Computers, Materials & Continua. 2021;68(2):2431–49.

56. Attrey A, Lesher M, Lomax C. The role of sandboxes in promoting flexibility and innovation in the digital age. Going Digital Toolkit Note, No. 2. Paris: Organisation for Economic Co-operation and Development; 2020. (https://goingdigital.oecd.org/data/notes/No2_ToolkitNote_Sandboxes.pdf, accessed 25 July 2023

57. Madiega T, Van De Pol AL. Artificial intelligence act and regulatory sandboxes. European Parliamentary Research Service, June 2022., PE 733.544; (https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf, accessed 25 July 2023).

58. First regulatory sandbox on artificial intelligence presented. European Parliamentary Research Service, June 2022. Brussels: European Commission (https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented, accessed 25 July 2023).

59. How should we engage and involve patients and the public in our work? London: Medicines and Healthcare products Regulatory Agency (MHRA); 2020 (https://www.gov.uk/government/consultations/how-should-we-engage-and-involve-patients-and-the-public-in-our-work, accessed 25 July 2023).

60. Stakeholder Engagement Framework. Canberra: Government of Australia, Department of Health and Aged Care; 2017 (https://www.health.gov.au/resources/publications/stakeholder-engagement-framework, accessed 25 July 2023).

61. Communication on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society. Brussels: European Commission; 2018 (https://digital-strategy.ec.europa.eu/en/library/communication-enabling-digital-transformation-health-and-care-digital-single-market-empowering, accessed 25 July 2023)

62. E-services. Singapore: Health Sciences Authority (HAS) (https://www.hsa.gov.sg/e-services, accessed 25 July 2023).

63. International Association for Public Participation (IAP2) Spectrum. International Association for Public Participation; 2007 (https://www.iap2.org/, accessed 20 March 2023).

64. Ho D. Artificial intelligence in cancer therapy. Science. 2020;367(6481):982–3. (https://science.sciencemag.org/content/367/6481/982, accessed 20 March 2023).

65. Ho D. Addressing COVID-19 drug development with artificial intelligence. Adv Intell Syst. 2020;2(5):2000070 (https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202000070, accessed 25 July 2023).

66. Blasiak A, Lim JJ, Seah SGK, Kee T, Remus A, Chye DH et al. IDentif.AI: Rapidly optimizing combination therapy design against severe acute respiratory syndrome Coronavirus 2 (SARS-Cov-2) with digital drug development. Bioeng Transl Med. 2020;6(1):e10196 (https://aiche.onlinelibrary.wiley.com/doi/10.1002/btm2.10196, accessed 25 July 2023).

67. Regulatory guidelines for software medical devices – a lifecycle approach. Singapore: Health Sciences Authority; 2019 (https://www.hsa.gov.sg/docs/default-source/announcements/regulatory-updates/regulatory-guidelines-for-software-medical-devices--a-lifecycle-approach.pdf, accessed 25 July 2023).

68. Ho D, Quake SR, McCabe ERB, Chng W J, Chow E K, Ding X et al. Enabling technologies for personalized and precision medicine. Trends Biotechnol. 2020;38(5):497–518. (https://www.cell.com/trends/biotechnology/fulltext/S0167-7799(19)30316-6, accessed 25 July 2023).

69. Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med. 2019;2:69. (www.nature.com/articles/s41746-019-0148-3, accessed 25 July 2023).

70. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40(8):577–91 (https://www.sciencedirect.com/science/article/pii/S0165614719301300#:~:text=AI%20techniques%20have%20advanced%20to,to%20assist%20human%20decision%2Dmakers.&text=We%20explain%20how%20recent%20advances,towards%20increasing%20trial%20success%20rates, accessed 25 July 2023).

71. Special access routes (medical devices). Import and supply of unregistered medical devices by request of qualified practitioners. Singapore: Health Sciences Authority; 2019 (https://www.hsa.gov.sg/medical-devices/registration/special-access-routes/qualified-practitioner-request, accessed 25 July 2023).

72. Complementary health products (CHP) classification tool. Singapore: Health Sciences Authority (https://www.hsa.gov.sg/CHP-classification-tool, accessed 25 July 2023).

73. Pagallo U, Casanovas P, Madelin R. The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the Web of Data. The Theory and Practice of Legislation. 2019;7(1):1–25.

74. Stakeholder communications. Sydney: Atlassian (https://www.atlassian.com/team-playbook/plays/stakeholder-communications-plan, accessed 25 July 2023).

75. Clarke R. Regulatory alternatives for AI. Computer Law & Security Review. 2019;35(4):398–409.

ANNEX. DEFINITIONS, FUNDAMENTAL CONCEPTS AND DECLARATIONS OF INTEREST

• Definitions and concepts

The FG-AI4H is proposing a new deliverable titled: "FG-AI4H terms and definitions" which aims to establish a new deliverable for the FG-AI4H with a glossary with agreed terminology in AI for health. The objectives of the new deliverable are the consistent use of terms across various deliverables, including WG-RC, and the promotion of harmonized use of important AI for health terms across the different disciplines involved in this cross-disciplinary field. However, this section applies to terms and concepts as they are used for the purpose of this document as part of the WG-RC. For more general terms across the FG, please refer to the FG-AI4H terms and definitions deliverable.

1. Artificial Intelligence

AI is a branch of computer science, statistics and engineering that uses algorithms or models to perform tasks and exhibit behaviours such as learning, making decisions and making predictions. The subset of AI known as ML allows computer algorithms to learn through data, without being explicitly programmed to perform a task *(1)*.

2. Trustworthiness

Trustworthy AI in the context of this document refers to AI systems and technologies that meet the stakeholder's expectation in terms of bias, explainability, provenance and other desirable characteristics. Therefore, stakeholders involved in the development, deployment or operation of such AI-based systems should be held accountable for their proper functioning.

3. Transparency

The term "transparency", in the context of this document, refers to issues such as sharing and making available to the appropriate entities the relevant plans, decisions and associated reasoning and the data/datasets utilized in the conception, development and ongoing deployment and monitoring of AI systems. Transparency is multifaceted and may include public dissemination by publications in peer-reviewed journals, and publishing and documenting pre-specifications for development processes, including clinical trials etc. Considerations should be given to factors such as data privacy and intellectual property, among others.

4. Documentation

For the purpose of this document, the term "documentation" refers to processes and methods used to document, retain and pre-specify critical development ideas, including the initial conception, validation, deployment and post-deployment plans – as well as relevant key decisions, choices and supporting rationale (e.g. selection of data/datasets) – used in the development of AI systems for health and therapeutic development throughout the total life cycle (e.g. from conception to post-deployment). Methods and approaches for risk and error management, reporting and detection of bias are all key areas for documentation. Documentation can also

help facilitate the understanding of the algorithm decision-making process (explainability). Documentation should allow for the tracing and audits of the development process and the steps taken in the development and validation of the AI system if needed and appropriate. This includes ensuring that changes and deviations from pre-specified approaches and protocols are tracked, recorded and justified. Although effective documentation is only one element that supports transparency, it is a key regulatory principle.

5. Privacy

Privacy is a broad and multidimensional concept. It is a universally accepted fundamental human right.[14] In nearly every nation, numerous statutes, constitutional rights and judicial decisions seek to protect privacy. The concept of privacy includes the control over personal information, often referred to as data or information privacy. Data privacy is focused on the use and governance of personal data, including implementing policies to ensure that consumers' personal information is being collected, shared and used in appropriate ways *(2).* Privacy risks include reidentification and the release of unwanted inferences about a data subject (e.g. whether they have a certain disease *(3).*

6. Data integrity

Data integrity can be defined as "the completeness, consistency, and accuracy of data"*(4).*

7. Data protection

Data protection is a more technical issue under the broader umbrella of privacy which includes more domains beyond the protection of an individual's personal data. However, for the context of this document, data protection includes the requirements and methods used to store and organize data in a physically secured manner to prevent unauthorized access and use. Data protection, although also a legal issue, is focused on securing data against malicious attacks and preventing the potential exploitation of stolen data for profit. While security is necessary for protecting data, it may not be sufficient for addressing privacy *(2)*.

8. Health data

Health data is personal data relating to a person's physical or mental health, and includes the provision of health-care services and information regarding a person's health status *(5)*. Health data are often considered to be a special category of personal data, or "sensitive" personal data, because of the nature and influence such data has on human lives and the impact on their fundamental rights and freedoms.

9. Sources of health data

Sources of health data include data acquired from digital health and medical technologies *(6)*, such as: wearable devices, digital health (or electronic health) applications, and medical devices and sensors; electronic health records and administrative hospital data; data from aggregated clinical trials; bioimaging and genomic data from the sequencing of human biological materials; health-related geospatial and contact-tracing data; insurance claims; and data from social media, smartphones and other electronic devices. The health data, or special personal data, derived from these sources, including heart rate, blood glucose, genetic predispositions, fitness levels, age, weight and so on, may be subject to data protection and privacy laws. Although these laws may vary from country to country, they will inform how the data are processed and for what purpose.

[14] According to the United Nations Universal Declaration of Human Rights of 1948, "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation."

10. Software as a medical device (SaMD)

SaMD is defined by the IMDRF as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device"*(7).*

11. AI system

The IMDRF *(1)* defines an AI system as a software that is developed with one or more of the techniques and approaches listed below* and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions that influence the environments they interact with.

*AI techniques and approaches:

(a) machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods, including deep learning;

(b) logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) statistical approaches, Bayesian estimation, search and optimization methods.

12. AI technology

In the context of this publication, the term "AI technology" refers to any AI technology (e.g. machine learning, deep learning, natural language processing, computer vision etc.) that is used to develop an AI system.

• Assessment and management of declarations of interest

All external experts submitted to WHO a declaration of interest disclosing potential conflicts of interest that might affect, or might reasonably be perceived to affect, their objectivity and independence in relation to the subject matter of the first meeting. WHO reviewed each of the declarations and found that four external experts declared interests in the topic under consideration; consequently WHO concluded to exclude those experts from contributing to the discussions on these subjects at the meetings and from contributing to the guidance. All remaining external experts were invited to participate in the discussions and contribute to the guidance. All experts participated in their individual capacities and not as representatives of their countries, governments or organizations. Therefore, the regulatory considerations in this guidance are not inclusive and regulatory bodies may have additional or different approaches.

References

1. Machine learning-enabled medical devices: a subset of AI-enabled medical devices: key terms and definitions. Proposed document posted for public consultation, 16 September 2021. International Medical Device Regulators Forum; 2021. (https://www.imdrf.org/sites/default/files/2021-10/Machine%20Learning-enabled%20Medical%20Devices%20-%20A%20subset%20of%20Artificial%20Intelligence-enabled%20Medical%20Devices%20-%20Key%20Terms%20and%20Definitions.pdf, accessed 25 July 2023).

2. What is privacy? International Association of Privacy Professionals (IAPP); 2020 (https://iapp.org/about/what-is-privacy/, accessed 25 July 2023).

3. Kearns M, Roth A. The ethical algorithm: the science of socially aware algorithm design. New York (NY): Oxford University Press, 2019.

4. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Draft guidance for industry. Silver Spring (MD): US Food and Drug Administration; 2021 (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory, accessed 25 July 2023).

5. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

6. Vayena E, Dzenowagis J, Brownstein JS, Sheikh A. Policy implications of big data in the health sector. Bull World Health Organ. 2018;96(1):66–8. doi:10.2471/BLT.17.197426. (https://pubmed.ncbi.nlm.nih.gov/29403102/, accessed 25 July 2023)

7. Software as a Medical Device (SaMD): Key definitions. International Medical Device Regulators Forum; 2013. (http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf, accessed 25 July 2023).

Digital Health and Innovation
Division of the Chief Scientist

World Health Organization
Avenue Appia 20
1121 Geneva 27
Switzerland

