WHO Consultation
Towards the
Development of guidance on ethics and governance of artificial intelligence for health

Geneva, Switzerland, 2–4 October 2019

Meeting report
WHO consultation towards the development of guidance on ethics and governance of artificial intelligence for health: meeting report, Geneva, Switzerland, 2–4 October 2019
ISBN 978-92-4-001276-9 (print version)

© World Health Organization 2021

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/).


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 http://www.who.int/about/licensing.

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. This publication contains the report of the “WHO consultation towards the development of guidance on ethics and governance of artificial intelligence for health” and does not necessarily represent the decisions or policies of WHO.
Contents

Acknowledgements iv
Executive Summary v

1. Introduction 1
2. The importance of ethics in artificial intelligence 2
3. Artificial intelligence is changing the health sector 3
4. Ethical use of artificial intelligence: principles, guidelines, frameworks and human rights standards 8
5. Challenges to ethics and governance 11
6. Addressing challenges to ethics and governance 16
7. Consensus points 20
8. Conclusion 20
References 22

Annex 1. List of participants 25
Annex 2. Agenda 28

Abbreviations and acronyms
AI artificial intelligence
LMIC low- and middle-income countries
OECD Organization for Economic Co-operation and Development
Acknowledgements

WHO convened an expert consultation to discuss ethics and governance in use of artificial intelligence for health, at WHO headquarters in Geneva, Switzerland, on 2–4 October 2019. The meeting was organized by Andreas Reis (Global Health Ethics unit) and Sameer Pujari (Department of Digital Health), under the guidance of Vasee Moorthy and John Reeder (Director, Department of Research for Health), Bernardo Mariano (Director, Department of Digital Health), and Soumya Swaminathan (Chief Scientist). WHO is grateful to the panel of experts who contributed to this report (Annex 1) and notably to the Chairs of the expert group, Professor Effy Vayena and Professor Partha Majumder, for their input in the conceptualization and realization of the project.

Special thanks are due to Tracy Swan, who served as rapporteur and prepared the first draft of this report. All the experts were invited to comment on the draft. Rohit Malpani, a consultant with the WHO Health Ethics and Governance Unit, supported the meeting, prepared the concept note and communicated with the invited experts.
Executive Summary

Artificial Intelligence (AI) consists of computing technology that performs similarly to human intelligence. AI might transform medicine, and its applications could be adapted for broader use, including pharmaceutical research and development, public health surveillance and health care management. While AI offers possibilities for improving health services and delivery, there are numerous questions about the ethics of developing and using these technologies, including whether and how AI can cross the digital divide to ensure that low- and middle-income countries (LMIC) will benefit from it. Many trans-national ethical, legal and social concerns should also be addressed, including equitable access, privacy, appropriate use and users, liability, bias and inclusiveness.

Although some government agencies, academic institutions, nongovernmental organizations and national ethics committees are addressing the ethical issues associated with use of digital technology in health care, there is no international guidance specific to health care. Harmonized ethics guidance is essential for the design and implementation of ethical AI in global health.

As the specialized United Nations agency for health, WHO is well-placed to develop a global framework for ethics and governance in AI to ensure that such technology is aligned with the broader aims of promoting fair, equitable global health, that it satisfies human rights standards and contributes to the commitments made by Member States to achieve universal health coverage. WHO is convening a group of international experts to develop guidance for Member States on ethics and governance of AI in health. The guidance development is a collaborative initiative of the Global Health Ethics unit and the Digital Health Department in WHO’s Division of the Chief Scientist and WHO collaborating centres and colleagues. The process of guidance development will be in line with WHO’s policies for developing norms and standards and will include several consultations of the established multi-disciplinary expert group, wide consultation with all relevant stakeholders and peer-review.

The present document is the report of the first expert meeting, which was held 2–4 October 2019 at WHO in Geneva. Participants discussed current and potential AI applications for global health in the context of existing frameworks and human rights standards. The participants considered the main challenges of ethics, governance and equitable access and discussed potential solutions. The expert meeting has established a foundation for the second meeting which will be held on 5 and 6 March 2020 in Copenhagen, Denmark. The publication of the WHO guidance document on Ethics and governance of AI for health is envisaged for completion in 2020.
1. Introduction

Digital technologies, machine learning and artificial intelligence (AI) could transform the fields of medicine, research and public health. The tools, methods and technology used for “Big Data” and in AI are being applied to health services and systems globally. While they hold great promise, this rapidly developing field raises trans-national ethical, legal and social concern about equitable access, privacy, appropriate uses and users, liability, bias and inclusiveness. Machine learning algorithms also pose novel ethical challenges in software engineering. Many unanswered questions remain about the ethical development and use of these technologies, including whether and how low- and middle-income countries (LMIC) will benefit from the developments.

Currently, there is no universal definition of AI. The term is commonly used to describe computing technology that works similarly to human intelligence, such as visual recognition, learning, reasoning, problem-solving, decision-making and adapting to change. The main objectives of AI technologies and applications are to emulate, complement, augment and enhance human capabilities in solving tasks. Although these processes may be like those performed by humans, their scope in AI systems is strongly limited, and they are not combined in the complex ways that allow human decision-making. Nonetheless, AI-based applications could be adapted for use in the health sector in research and development, public health and health care management. Currently, AI-based applications in health care are used mainly in high-income countries for individualized health care and for strategies such as automated diagnostics and promoting healthy behaviour.

Several government agencies, academic institutions, nongovernmental organizations and national ethics committees have begun to address the ethical issues and challenges associated with digital technology in general, but there is no international guidance on its use in health. Harmonized ethics guidance should therefore be developed for the design and implementation of AI in global health. For this reason, WHO has started an initiative to develop a global framework for ethics and governance to ensure that AI-based technologies are consistent with the broader aim of promoting fair, equitable global health and assuring universal health coverage.

Building on a previous WHO consultation on ethics, Big Data and AI (1), WHO aims to prepare guidance for Member States on ethics and governance of AI in health in 2020. This initiative is a collaboration of the Global Health Ethics team and the Digital Health Department in WHO’s Division of the Chief Scientist, other WHO colleagues, and WHO collaborating centres with specialized expertise. A group of external international experts was convened to advise WHO on the development of a globally relevant framework and guidance.

WHO’s Department for Digital Health and Innovation

WHO has established a new Department of Digital Health and Innovation to enhance its role in assessing digital technologies and supporting Member States in prioritizing, integrating and regulating these technologies. WHO is committed to harnessing the possibilities of digital technologies through its programmes and collaboration and by:

- issuing evidence-based guidelines on mobile and digital technologies and recommendations after extensive consultation with experts on ways to maximize the impact of these tools on health systems and people’s health (2);
- communicating with other stakeholders to ensure that new and existing digital platforms have a positive impact on people’s lives; and
• participating in new collaborations and partnerships, such as the Precision Public Health Initiative (see below), Be He@lthy, Be Mobile (3), a joint initiative with the International Telecommunication Union on use of mobile technology to improve the health of people at risk of or with a noncommunicable disease and the WHO–International Telecommunication Union and European Commission mHealth Innovation and Knowledge Hub (4), a collaboration funded by the European Union Horizon 2020 to establish an mHealth Innovation and Knowledge Hub for Europe.

The mission of WHO’s Department for Digital Health and Innovation is to ensure that digital health contributes to attainment of the highest level of health for all through the “triple billion” goals of the WHO 13th General Programme of Work and Sustainable Development Goal 3, to ensure healthy lives and promote well-being for all people at all ages (5).

At the World Health Assembly in 2018, Member States unanimously adopted resolution WHA71.7 (6), which calls on WHO to prepare a global strategy on digital health to support national health systems in achieving universal health coverage. In September 2019, the United Nations hosted a high-level meeting to finalize commitments and recommendations for achievement of universal health coverage by 2030.

WHO guidance on ethics and governance of artificial intelligence in health

In 2019, WHO’s Digital Health Department and Global Health Ethics unit, under the Office of the Chief Scientist, initiated a project to prepare guidance on ethics and governance for the design and use of AI in global health. The first step was a joint international consultation in October 2019, with participants from government, inter-governmental agencies, academia, civil society and industry. The objectives of the consultation were to:

• discuss current and anticipated uses of AI technology for health and the ethical considerations and human rights principles that should guide their use;
• examine the initiatives of international organizations and governments; and
• give advice on the format and content of guidance on ethics and governance to be elaborated at further consultations planned for 2020.

This report summarizes the presentations and discussions at the meeting.

2. The importance of ethics in artificial intelligence

AI could help shape the future of individual and global public health and help countries to achieve universal health coverage. However, AI can be part of both the solution and the problem. The risks and challenges of AI in health care must be considered at the same time as its potential benefits.

AI could extend the scope, transparency, safety, effectiveness, reliability and accessibility of health services and information to reach more people, including marginalized and vulnerable populations. AI could improve public health surveillance, forecasting, modelling and monitoring, facilitate training of the health workforce, introduce innovation and efficiency into health systems, improve clinical outcomes, ensure cost-effective planning, accelerate and improve drug development and optimize operations and health care provision.

AI could also exacerbate the unequal distribution of health care, reduce the health care workforce, reduce the skills of health workers, perpetuate or increase disrespectful clinical interactions and widen differences in health care outcomes within and between countries. The benefits of technology are already distributed unevenly and unequally between and within rich
and poor countries. The digital divide could worsen inequitable access to health care, whether by geography, gender, age, access to devices and connectivity.

While privacy, confidentiality and informed consent are pillars of patient rights worldwide, these rights could be dramatically redefined or undermined when digital technologies take hold and expand. The performance of AI depends on the nature and extent of data. Use of restricted, poor and heterogeneous 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. It has been shown that commercial prediction algorithms can identify complex health needs, but they can also result in significant racial bias, so that black patients are at a greater disadvantage than white patients when health care costs are used to train the algorithm (7). Some AI algorithms are based on images of uneven quality and resolution, which could adversely affect the accuracy and reliability of the systems’s outputs. The prices of imaging devices, blood tests, electroencephalograms and other tests for health and disease assessment are highly variable. Less expensive devices, of compromised quality, are often the only ones available in LMICs, thereby increasing the digital divide.

Ethical application of AI could mitigate or eliminate some of these risks, by collection of appropriate, high-quality data, appropriate design of AI and AI software, policies and laws based on widely shared ethical principles or work by programmers, health care workers, patients and policy-makers to address harms and wrongs as they emerge. Even the best-designed systems will not increase the capacity of LMIC health systems if they are not incorporated into clinical contexts and if the savings and gains in efficiency from these systems are used solely to replace current capacity at a lower cost instead of increasing current capacity to provide better health services to more people at lower cost.

Similarly, widespread use of AI systems could compromise the integrity of LMIC health systems. While an AI system for reading a scan may increase the capacity of a health system to provide a diagnosis, maintenance of the system may require additional expenditure and personnel. If the necessary maintenance and infrastructure are not locally available, adoption of AI systems could render LMIC health systems more dependent on companies in high-income countries for infrastructure and service.

3. Artificial intelligence is changing the health sector

AI applications have advanced rapidly with improvements in computing power and analysis of larger volumes of data; however, when AI is used for patient care, lack of or inadequate computing power could delay inferences for treatment. At present, AI is confined mainly to “artificial narrow intelligence”, a form of AI in which a technology outperforms humans in a narrowly defined task. If AI evolves over the next two decades towards “artificial general intelligence” or even to “artificial super intelligence” (both of which are highly speculative), the potential of AI in health care could multiply.

AI will be mainstreamed into health care systems through “push factors” such as public, private and philanthropic investment. A recent example of expanding use of AI is the Precision Public Health Initiative (8), a data collaborative initiated by the Rockefeller Foundation, with WHO, UNICEF, global health funding agencies, ministries of health and technology companies, with initial funding of US$ 100 million. The aim is to extend use of AI and data science in LMICs by providing the latest technology to parts of the world that have not yet benefited.

Adoption of AI will also be driven by “pull factors”, such as the anticipated savings that AI could generate, as certain technologies are used more widely in health care and medicine.
Box 1. Top 10 artificial intelligence applications in health and related savings estimated by 2026 (9)

- Robot-assisted surgery: US$ 40 billion
- Virtual nursing assistants: US$ 20 billion
- Administrative workflow assistance: US$ 18 billion
- Fraud detection: US$ 17 billion
- Reduction of dosage error: US$ 16 billion
- Connected machines: US$ 14 billion
- Clinical trial participant identifier: US$ 13 billion
- Preliminary diagnoses: US$ 5 billion
- Automated image diagnoses: US$ 3 billion
- Cybersecurity: US$ 2 billion
- Total: ~US$ 150 billion by 2026

Below, we examine how health care supported by AI will evolve, including in diagnosis, clinical care, public health surveillance, drug discovery and development, and the allocation and training of human resources for health. Numerous obstacles should also be kept in mind. The promises of AI have until now exceeded the underlying science and the readiness for implementation in patient care (10).

*From diagnosis to prediction with artificial intelligence*

AI has been shown to be effective for specific pattern recognition and tasks in radiology and medical imaging. Such applications, however, are not currently in wide use, and AI is not yet used in routine clinical decision support. Currently, AI is being evaluated for diagnosis in oncology in radiological applications (thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging and dose optimization for radiology treatment), for non-radiological applications (dermatology, pathology) and for RNA and DNA sequencing to guide immunotherapy. Nevertheless, few such systems have been evaluated in prospective clinical trials, and a recent comparison of the ability of deep-learning algorithms and of health care professionals to detect diseases from medical imaging found that AI is equivalent to human medical judgement in specific domains and applications in specific contexts (11).

While AI is used for disease prediction and diagnosis, it could also be used to assess the relative risk of disease, which could be used for prevention of lifestyle diseases, such as heart disease and diabetes. As AI improves, it could initially assist medical providers in making faster diagnoses. It could also be used for prompt detection of stroke; in detecting pneumonia, breast cancer and other diseases by imaging (12,13); in echocardiography for diagnosing coronary heart disease (14) and in screening to predict psychotic episodes by analysis of speech patterns (15). AI could assist health care providers in predicting illness or major health events before they occur. For example, early studies with limited datasets indicate that AI could be used to diagnose Alzheimer disease years before symptoms appear (16).

Box 2: Artificial intelligence for health in LMIC

In LMIC, AI could be used for low-cost diagnostics, especially where trained staff are scarce or non-existent. Telehealth and chatbots could provide channels of direct care, and mHealth technology could
be used to collect health information in communities, provide diagnoses, deliver clinical information to health care providers and train unskilled workers.

Several AI applications are being developed or used in LMIC. They include point-of-care detection of cervical cancer (17) and software to create an encephalogram in 6 min in India (18); Babyl, a Rwandan chatbot and telehealth system that directs people to medical personnel when indicated for appointments, prescriptions, laboratory tests and delivery of results (19); a Ugandan app that helps to diagnose malaria (20); a Pakistani app that locates nearby pharmacies, facilitates prescription delivery and supports adherence to medicines (21); and mHealth technology in Nigeria to monitor vital signs in pregnant women and young children (22) and an app that can detect asphyxia by analysing infant cries (23).

While AI-based diagnosis is near-term and its efficiency can be tested, thereby mitigating potential harm, efficacy and accuracy may be more difficult or impossible to achieve in long-term predictions. The risk of harm is therefore increased dramatically, as predictions could affect an individual's health and well-being or restrict or unnecessarily expend scarce resources. Sharing predictions with people who did not consent to surveillance or detection also raises a serious ethical concern with respect to informed consent and individual autonomy.

The accuracy of diagnosis and prediction with AI depends on preconditions such as expensive, diverse, annotated input data (including data on environmental exposure), and algorithms designed to avoid bias and judicious use. Algorithms for diagnosis and prediction can be harmful if they are weak because of inadequate, poor-quality or non-representative data that do not adequately cover all sections of the population. AI may also be inefficient if it is not designed properly, applies only under certain assumptions, is based on biased data or is applied out of context. The predictions of current AI systems are based on large numbers of associations among variables that may not be directly causally involved in the disease that is being predicted. Thus, changes in the way data are collected, in social conditions or in the behaviour of patients in response to use of such systems can degrade or alter their performance in ways that might be difficult to detect in real-world settings.

*Evolution of artificial intelligence-guided clinical care*

Use of AI in medicine was first mentioned 60 years ago (24). The objective of AI was to help junior clinicians make accurate diagnoses, especially in settings where clinical expertise was not easily accessible. AI solutions included diagnosis, identification of risk factors for disease progression and the best course of treatment, given interacting factors such as cost, effectiveness and risks. The complexity and potential richness of AI in medicine inspired technologies that are still used today.

Efforts are being made to make AI systems “human-centric”, and models are being developed to work with and for people, and to become human–machine teams that work effectively together. AI technologies are being changed from clinician-centric to patient-centric, with the objective of facilitating shared decision-making with patients, their caregivers and their communities. They are also being changed from evidence-based medicine to precision medicine and ultimately to personalized medicine. Other work is being conducted on AI that can expect and react to biases and emotions in human beings.

At present, clinical experience and knowledge about patients remain necessary, as AI is not a substitute for clinical due diligence. Clinicians use AI for decision support, to reduce their cognitive load, integrate patient records during consultations, identify patients at risk and vulnerable groups and to aid difficult diagnostic or treatment decisions. The current AI
technology that holds the most promise in clinical care includes tools for patients, lay workers and clinicians, and robotics for certain clinical tasks. Self-help AI provides patients with rapid answers and guides them to services, while digital, environmental and wearable devices allow health profiling, health coaching, reminders and communication between caregivers and clinicians.

In time, AI is expected to assist in the provision of clinical care and assume other human roles. For example, AI could evolve to address scientific, clinical, social and economic aspects of dementia, which will be relevant for ageing patients and their caregivers. Diagnosis of dementia is challenging, as Alzheimer’s and other neurodegenerative diseases are complex, with unknown causes and some signs, signals and biomarkers that vary at different stages. AI could be used to analyse symptoms and the results of psychometric tasks and neuroimaging for early diagnosis, thus enabling preventive treatment. AI could also ease the burden of caregiving for dementia patients, with monitoring and reminder devices for patients and communication with their caregivers.

At present, although AI can solve certain problems and perform certain tasks, it has not reached its full potential. Challenges to the progress and deployment of AI in medicine include interactions of social and medical issues and the differing skills, preferences, biases and use patterns of those who use it.

At the consultation, panellists discussed the ethical issues related to interventions with machines in the absence of physicians and ensuring use of AI to benefit people without compromising the quality of health care. There are technological challenges for wider use of AI in medicine. Although many prototypes developed in both the public and private sectors performed well in field tests, the technologies cannot be translated, commercialized or deployed. An additional obstacle is constant changes in computing and information technology management, whereby systems become obsolete and companies disappear.

Health care workers will have to significantly adapt their clinical practice as use of AI increases. Therefore, clinicians who understand the clinical workflow should be involved in building AI tools from the beginning, so that health services feel “ownership” of the tools, and so that they are patient-centred, enhance clinician–patient relationships and focus on what they are suited for.

Putting diagnostic tools into the hands of lay workers through AI changes the way people think about the role and function of doctors. AI could automate tasks, so that doctors have time to listen and address fears and concerns and ask about unrelated social factors, although clinicians may still worry about responsibility and accountability and inaccurate test results. Even if technology makes such gains, they will materialize only if the individuals who manage health systems use them to extend the capability of health systems in other areas. Doctors will need new skills to communicate risks, predictions and trade-offs to patients and to understand AI technology. Such shifts in the provision of clinical care may shift legal accountability from doctors and scientists to engineers and coders.

From artificial intelligence-enabled public health surveillance to refined prediction

To date, public health surveillance has been based on collecting evidence and using it to create mathematical models to make decisions. Technology has revolutionized 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) that can be applied to public health. Videos (e.g. YouTube) are a rich source of information that can be transformed into health insights. Public health institutions have not yet made full use of these sources of data. Google Flu Trends (25), for example, is based on search engine queries about
complications, remedies, symptoms and antiviral medications for influenza, which are used to estimate and predict influenza activity.

Data are useful only when appropriate models are used. Similarly, machine-learning and algorithms may be more valuable and effective when augmented by digital traces of human activity. Models have evolved from mechanistic models, such as use of data from air travel networks to predict the possible emergence of pandemics, to “black box” models with unknown components and ingredients. These advances have enabled a more accurate model of human population behaviour, with greater detail, complexity and capability for forecasting.

Thus, surveillance is already 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 (26). Similarly, Microsoft researchers have found early evidence of adverse drug reactions from web logs. 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 (27).

WHO is developing EPI-BRAIN, a global platform that will allow experts in data and public health to analyse large datasets for emergency preparedness and response. EPI-BRAIN is intended to harness Big Data and AI to mitigate the impact of epidemics. It will enable forecasting and early detection of infectious threats and their impact from scenarios, simulation exercises and sharing of insights to improve coordinated decision-making and response.

*Artificial intelligence for drug discovery and development*

Drug discovery and development is lengthy: development of a new drug often takes a decade. 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 its development to reduce the delay from decades to months or even weeks in the next 20 years. As an example, 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 a safe, effective therapy (28).

At present, there is collaboration between humans and machines, led either by humans or by AI with human oversight. In the next two decades, as work with machines is optimized, AI could evolve. Computing could allow 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 Big Health Data to identify structures and features, while recreating the body and its organs on AI chips. By 2040, some trials might be virtual – without animals or humans – based on models of the human body, tumours, safety, efficacy, epigenetics and other parameters. Prescription drugs could be designed for each person.

For this to occur, challenges in data harmonization, sharing and interoperability will have to be addressed, as well as ethical dilemmas arising for other uses of AI in health care, such as algorithmic bias and a general public demand for “explainability”, which will affect the use of AI in drug development (29). Issues of ownership and liability may arise. The benefits include more rapid marketing of medicines (which may also be a risk) and possible use of such approaches in antimicrobial resistance and rare diseases.
4. Ethical use of artificial intelligence: principles, guidelines, frameworks and human rights standards

The benefits of AI come with technical, ethical and governance challenges. Efforts to address these challenges have been fragmented and limited. Increasing numbers of principles and guidelines have been developed for different applications of “ethical” AI in the private and public sectors and in research institutions; however, there is no consensus on its definition, best practices or ethical requirements, and different legal regimes and governance models have been associated with each set of principles. The participants in the consultation discussed the ethical aspects of the goals and objectives of corporate and non-corporate actors in the health sector. They noted that most reports are from high-income countries, produced by private companies, government agencies, academic and research institutions, intragovernmental organizations, non-profit organizations and scientific and professional societies (30).

General principles of artificial intelligence

One review of more than 50 proposals for general AI principles (31) was based upon identifying 10 principles: humanity, collaboration, sharing, fairness, transparency, privacy, security, safety, accountability and long-term AI. None of the proposals included more than two thirds of the principles. Humanity, fairness, privacy and safety were most cited, while references to long-term AI, collaboration and sharing were less frequent. In another mapping and analysis of current principles and guidelines for ethical use of AI, Jobin et al. (30) noted convergence on transparency, justice, fairness, non-maleficence and responsibility, while other principles such as privacy, solidarity, human dignity and sustainability were under-represented. There was some disagreement about the interpretation, domain, stakeholders and implementation of guiding principles.

Several inter-governmental organizations and countries have proposed such principles (Box 3).

<table>
<thead>
<tr>
<th>Box 3: Examples of artificial intelligence principles proposed by inter-governmental organizations and countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The Organization for Economic Co-operation and Development (OECD) Recommendation of the Council on Artificial Intelligence (32), the first intergovernmental standard on AI, was adopted in May 2019 by OECD’s 36 member countries, along with and a range of partner economies which have since adhered to them. The OECD AI Principles (33) also provided the basis for the G20 AI Principles endorsed by Leaders in June 2019. While not legally binding, OECD recommendations do carry a political commitment, and in other policy areas (e.g. privacy and data protection) have proved highly influential in setting international standards and helping governments to design national legislation. The OECD also will launch on 27 February 2020 an online platform for public policy on AI, the AI Policy Observatory. (34) The Observatory combines resources from across the OECD with those of partners from all stakeholder groups to provide a comprehensive database of AI policies from around the world, trends and data as well as multidisciplinary, evidence-based policy analysis on AI, and facilitate multi-stakeholder dialogue. The OECD is cooperating closely on this and other initiatives on the ethical implications of AI with the Council of Europe and UNESCO.</td>
</tr>
<tr>
<td>• 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 (35).</td>
</tr>
</tbody>
</table>
• The European Commission has appointed 52 representatives from academia, civil society and industry to its High-level Expert Group on Artificial Intelligence and issued Ethics Guidelines for Trustworthy Artificial Intelligence (36). Pending completion of a pilot project, the revised document is expected in 2020.
• China has issued the Beijing Declaration on Artificial Intelligence, company principles (such as those of Baidu and Tencent) and principles prepared by academic institutions, such as the Tsinghua University Centre for International Security and Strategy.
• In Singapore, a series of AI governance and ethics initiatives was designed to build an ecosystem of trust to support AI adoption. 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; and a research programme on the governance of AI and data use established in partnership with the Singapore Management University in September 2018 (37).

Artificial intelligence principles for health

No principles for use of AI in health have yet been proposed. Before WHO’s work on guidance for the ethics and governance of AI in health, the WHO Global Conference on Primary Health Care issued the Astana Declaration (38), 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 recommended development of guidance and principles for the use of AI in general and as related to 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 2017 International Bioethics Committee report on Big Data and Health (39), which identified important elements of a governance framework; the 2017 World Commission on the Ethics of Scientific Knowledge and Technology Report on Robotics Ethics (40); the 2019 Preliminary Study on the Ethics of Artificial Intelligence (41), which raised ethical concern about education, science and gender; the Recommendation on Ethics of AI to be considered by UNESCO’s General Conference in 2021; and the World Commission on the Ethics of Scientific Knowledge and Technology Report on the Internet of Things.

In 2019, the United Kingdom of Great Britain and Northern Ireland’s National Health Service released a code of conduct with 10 principles for the development and use of safe, ethical, effective data-driven health and care technologies (42).

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 is working between October 2019 and December 2021.

Artificial intelligence and human rights

Machine learning systems could advance human rights (including the human right to health), yet could undermine core human rights standards. Human rights law and standards protect individuals against discrimination, promote inclusiveness, diversity and equity and safeguard
equality. Human rights relevant to AI also include privacy, data protection, freedom of expression, equality before the law and equitable access to essential services (including health care).

Human rights organizations have adapted existing human rights laws and standards to AI and are reviewing them in the face of challenges posed by AI. The Council of Europe is examining the feasibility and potential elements of a legal framework for the development, design and application of digital technologies, based on its standards on human rights, democracy and the rule of law. The Toronto Declaration (43) 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, protects people from discrimination and promotes inclusion, diversity and equity while safeguarding equality and effective redress and remedy.

Legal frameworks for human rights, bioethics and privacy that have been adopted by countries are applicable to several aspects of AI in health. They include Article 8 of the European Convention on Human Rights: the right to respect for private and family life, home and correspondence (44); the Oviedo Convention on Human Rights and Biomedicine (45), which covers ethical principles of individual human rights and responsibilities; Convention 108+ for the Protection of Individuals with Regard to Automatic Processing of Personal Data (46) 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+ (47).

Yet, even with robust human rights standards, organizations and institutions recognize that better definition is required of how human rights standards relate and apply to AI, and that new laws and jurisprudence are required to address the intersection of AI and human rights. New legal guidance has been prepared by the Council of Europe. In 2019–2020, the Council set up the Ad-hoc Committee on Artificial Intelligence to conduct broad multi-stakeholder consultations 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 (48), also based on the protection of human dignity and safeguarding human rights and fundamental freedom.

---

**Box 4. Council of Europe Guidelines on Artificial Intelligence and Data Protection**

1. The protection of human dignity and safeguarding of human rights and fundamental freedoms, in particular the right to the protection of personal data, are essential when developing and adopting AI applications that may have consequences on individuals and society. This is especially important when AI applications are used in decision-making processes.

2. AI development relying on the processing of personal data should be based on the principles of Convention 108+. The key elements of this approach are: lawfulness, fairness, purpose specification, proportionality of data processing, privacy-by-design and by default, responsibility and demonstration of compliance (accountability), transparency, data security and risk management.

3. An approach focused on avoiding and mitigating the potential risks of processing personal data is a necessary element of responsible innovation in the field of AI.

4. In line with the guidance on risk assessment provided in the Guidelines on Big Data adopted by the Committee of Convention 108 in 2017, a wider view of the possible
outcomes of data processing should be adopted. This view should consider not only human rights and fundamental freedoms but also the functioning of democracies and social and ethical values.

5. AI applications must at all times fully respect the rights of data subjects, in particular in light of article 9 of Convention 108+.

6. AI applications should allow meaningful control by data subjects over the data processing and related effects on individuals and on society.

In addition, the Council of Europe’s European Commission for Efficiency of Justice has an ethical charter that includes five principles relevant to use of AI in health.

Although human rights norms and principles may guide laws and policies in many countries, some countries do not fully apply international human rights standards in formulating or enforcing their laws and policies.

Role of WHO

As the specialized United Nations agency for health, WHO is preparing an ethics and governance framework for its Member States to ensure that AI promotes fair, equitable global health. Its leadership will allow the global community to recognize possibilities for synergy and streamlining of initiatives for equitable use of AI for health.

5. Challenges to ethics and governance

The participants in the consultation considered several challenges to ethics and governance, including:

- ensuring equitable access to AI and determining how AI is affected by, and affects the digital divide;
- preserving individual rights of autonomy, privacy, informed consent and freedom from bias and discrimination;
- ensuring understanding of how AI functions and makes decisions (transparency and explainability);
- ensuring equal access to databases for all users and not only those with more capacity to pay for access;
- preserving human control of AI; and
- strengthening public oversight and regulation of the private sector.

Equitable access to artificial intelligence and the digital divide

Although the cost of digital technologies is falling, access has not become more equitable. The “digital divide” refers to uneven distribution of access to, use of or effect of information and communication technologies among any number of distinct groups. For example, 1.2 billion women (327 million fewer women than men) in LMICs do not use mobile Internet services because they cannot afford or do not trust the technology. Gender is only one dimension of the digital divide; others include geography, culture, religion and language and among generations. The digital divide begets other disparities and challenges, many of which will affect the use of AI and how AI can itself reinforce and exacerbate the disparities.
The human and technical resources required to fully realize the benefits of digital technologies are also unequally distributed, and infrastructure to operate digital technologies may be limited or non-existent, limiting the collection, analysis, management, storage, transfer and use of data. Data may be difficult to collect because of language barriers, and mistrust may lead people to provide incorrect or incomplete information. Often, irrelevant data are gathered, which undermine the overall quality of a dataset. Data storage and handling are costly and may require significant maintenance skills, which might not be available. Simple processes and functions that are often taken for granted in high-income countries, such as data transfer, may present extraordinary challenges. For example, the time required to transfer a genomics dataset, which may be 7–20 terabytes, would be 7 days in places where the speed to copy is 25 megabytes per second, with an additional 7 days required for encryption and 6 more days for the recipient of the encrypted data to delete them after decryption. In many parts of the world, it is difficult to ensure uninterrupted Internet connectivity, and the whole process might have to be restarted.

AI could improve access to better diagnostics and therapies and reduce disparities in access to health care. This value proposition depends on overcoming the digital divide. Otherwise, the divide will only grow within and between countries, with social and ethnic disparities. The divide will be due to a lack of infrastructure and resources and also to differences in income, education and understanding and support for AI in the general public. Self-care tools that could reduce barriers to health care will be only be equitably accessible if there is widespread digital literacy and access to technology.

AI that is designed for worldwide use by companies and individuals in certain countries may lead to racial and gender disparities among technology and AI developers and in AI technologies based only on data for certain populations. They might therefore be irrelevant at best or discriminatory, biased or inaccurate at worst. In the relentless push for optimization, AI might ultimately exclude or fail to account for the least privileged in society. The question is how to create AI in a way that benefits those who have had the least input into its creation.

Preserving individual rights, including the rights to privacy and confidentiality, autonomy and informed consent and freedom from bias and discrimination

The panellists at the consultation debated ways to make AI consistent with longstanding individual rights. As AI expands into health care, the exercise of those rights may be increasingly difficult to protect and enforce, especially as rapid diagnosis and epidemiological concerns move from human beings to prediction-based algorithms and technologies.

Privacy and confidentiality

Privacy and confidentiality are concepts that extend beyond data protection. In the era of Big Data and AI, privacy is also the right to be left alone, to not know one’s health status and to opt out of exposure to such technologies. However, there are, consequences for individuals, communities or groups that choose to opt out, whether due to precaution, disability, disease or misfortune, as exclusion creates bias. There is also a risk of re-identification of individuals who had assumed that their data had been “de-identified”.

One panellist commented that use of Big Data and AI need not increase the risks of people who are the sources of the data, and that nimble professional and social governance, combined with security measures, could ensure privacy and confidentiality. The panellist also noted that the benefits of public health and science may outweigh the right to privacy, insofar as safeguards
are in place. Many panellists considered that precedent and customary practice in other areas of public health and science were not applicable because the companies that design and manage AI technologies depend predominantly on surveillance of individuals and aggressive use of individual and population data to make technologies effective and use of “surplus” data for commercial purposes. For example, Eric Schmidt, a former Chief Executive Officer of Google, once remarked: “We know who you are, where you have been, more or less what you think.”

Freedom of choice, control and autonomy
Panellists questioned whether individuals could safeguard their autonomy, freedom, choice and control over their own health and well-being and about patient ownership of information. The advent of prediction-based technologies will challenge the ability of humans to exercise freedom and choice. AI and Big Data, combined with the science of “nudging”, could quickly transform a useful method to promote healthy behaviour into powerful technologies with an outsized impact on the choices and freedom that people expect to exercise in their daily lives. As AI can be used to predict disease (see above), questions arise about ensuring that individuals still have the right to make certain decisions or whether algorithm-based predictions will discourage or prevent certain behaviour. There is also concern that such technologies could nullify the right of an individual to consent to surveillance and diagnosis, as diagnosis increasingly predicts the onset of illness.

Informed consent
Panellists questioned whether true informed consent for the use of data is even possible, as opting out of providing data for such technologies can result in sub-optimal or even biased or discriminatory functioning. There is also concern about the practicality of informed consent, as individuals struggle to keep track of all the uses of their data and cannot provide informed consent for each use. As some uses of data are less well defined than others, free prior and informed consent may not be possible and must be supplemented with stronger governance of the use of data.
As AI is increasingly used to diagnose and also to recommend a course of treatment, there is a broader risk that delegation of medical decisions, whether by a practitioner or a patient, will be increasingly surrendered to an AI technology, especially if the technology is presented to the patient as providing far more insight than a physician into his or her health status and prognosis. Treatment decisions require baseline data based on diverse ethnicity, geography, gender, age and other variables; therefore, databases established to support AI decision-making must have adequate representation of different subgroups.

Freedom from bias and discrimination
Bias and discrimination can pervade many aspects of AI, and the digital divide contributes. As noted above, the likelihood that AI-based technologies are and will continue to be developed by one demographic group and gender increases the likelihood of certain biases in the design of the technology. There is also potential bias in data collection and integration when individuals and communities choose not to participate, resulting in lack of data on those groups. Lack of data adversely affects the performance of AI algorithms, as the inferences may not be generalizable. For example, data on certain population subsets may be difficult to collect if it depends on expensive devices such as wearable monitors.
Some technology groups and universities have attempted to remove bias from AI technologies; however, they have been only partially successful, as algorithmic finesse and computing power cannot squeeze out information that is not present. In other words, a model’s efficacy depends on the datasets used to “train” it and the domain knowledge, if applicable, that was incorporated. If the datasets and knowledge are not sufficiently broad and inclusive, the outcomes will also not be fully inclusive. There may be a tendency to a form of sampling bias, whereby ethnic minorities, older people, rural communities and other disadvantaged groups are excluded. Furthermore, biases in data due to poor clinical practice, such as disparities in health care for ethnic and racial minorities, can be reproduced in AI technology.

Even AI that is effective for diagnoses or for accurate prediction of disease may present significant risks of bias and discrimination against individuals because of existing conditions or their predisposition to health conditions. Such bias may manifest itself in the workplace, in health insurance or in access to health care resources, benefits and other opportunities. AI could also, at a population level, 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.

Knowledge (transparency and explainability) of how artificial intelligence functions

Physicians and other medical professionals have a duty to ensure collaborative decision-making with patients. To do this, they must help patients understand complex medical information so that they can make decisions that reflect their considered values. There is widespread concern that the “black box” nature of many AI systems will undermine this duty, with fear of “unaccountability and domination by systems that arbitrarily restrict stakeholder autonomy and represent a conduit for experts to covertly impose arbitrary preferences on stakeholders” (49).

Knowledge is predicated upon transparency and explainability. 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 even their own developers do not fully understand. Concern about lack of transparency and explainability is heightened in communities that are “on the wrong side” of the digital divide. It also undermines the obligation of medical practitioners to fully understand why certain decisions are made and to communicate the explanations to their patients, who may be uncertain about how confident they should be in a diagnosis by AI. It also raises fears of “unaccountability and domination by systems that arbitrarily restrict stakeholder autonomy and represent a conduit for experts to covertly impose arbitrary preferences on stakeholders” (49).

Although transparency and explainability are often considered important for promoting accountability in the use of AI systems, it is often unclear what must be transparent and what must be explainable. Informed consent and warranted trust in AI systems depend on transparency about the goals advanced with AI systems. Clinicians must be able to explain to patients the clinical goals being advanced by a system, e.g. “the system takes an image of your retina and scans it to detect diabetic retinopathy”. Without the system, patients would be asked to submit to arbitrary testing or procedures that are not clearly linked to the goals of care.

It has also been argued that AI systems must be transparent in a deeper sense, namely, that clinicians must be able to explain how the system arrived at a diagnosis so that they can audit the decision. Further, some argue that, if a trade-off must be made between such transparency and accuracy, transparency should be preferred. This requirement reaches beyond what may be possible, or even desirable, in the medical context. While it is often possible to explain to patients why a 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 for a particular purpose before their mode of action is understood.

Trust in decisions and expert recommendations depends on the ability of experts to explain why
a system is the best option for achieving a clinical goal. Explanations of this kind should be
based on reliable evidence on the superior accuracy and precision of AI systems over
alternatives. The evidence should be generated by prospective testing of the system in
randomized trials and not the performance of the systems against existing datasets in a
laboratory.

Understanding how a system arrives at the judgements it makes may be valuable for a variety
of reasons, but it should not take precedence over or replace sound, prospective evidence of
that system’s performance in prospective clinical trials. Explanations of how a system arrived 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.

In summary, providers must be transparent about the clinical goals that an AI system is
supposed to advance, and their decision to use such a system to advance that goal must be
based on sound empirical evidence of the accuracy and precision of the system in prospective,
randomized studies. They must also be able to explain why they consider that they are using
the system under conditions that preserve the accuracy and precision of its results and its
merits and risks relative to alternative methods of reaching the same clinical goal.

Control and oversight of artificial intelligence

AI applications raise strong concern about whether humans will not only understand why an
algorithm decides but whether medical decision-making will be surrendered to AI. In some
situations, medical providers may demonstrate “automation bias” and not consider whether an
automated technology meets their needs or those of the patient. As autonomous systems such
as driving and warfare, have emerged, there has been growing concern about whether humans
can exert “meaningful control” over such technologies or whether the technologies will
increasingly make decisions independently of human input and control.

Use of AI systems to make specific, well-defined decisions may be entirely justified, if there is
compelling clinical evidence that the system performs that task better than a human. Leaving
decisions to humans when machines can perform them more rapidly, accurately and with
greater sensitivity can mean that some patients suffer avoidable morbidity and mortality without
the prospect of some offsetting benefit.

In some cases, automation of routine, mundane functions, such as recording information, will
actually liberate a medical provider to exercise the art of healing by being present, listening,
building a relationship with a patient, exploring the patient’s worries and fears and asking about
other factors. Freedom to focus on patients can enhance clinician–patient relationships while AI-
guided machines automate certain aspects of caregiving.

Panellists expressed broader concern about the increasing extent to which corporations exert
power over the development, deployment and use of AI in health care (including drug
development) and the extent to which corporations exert power and influence over individuals
and governments. This is a concern, as much of the data, computing power, human resources
and technology are increasingly concentrated in a few companies.
Panellists expressed concern that, without a strong government role, corporations might ignore the needs of citizens, particularly of those at the margins of their societies and the global economy, or that some countries and communities would be used for experimentation rather than healthy investment and collaboration.

Oversight by governments might also be integrated into public–private partnerships, whereby a collaboration is used for purposes and motives that are not primarily intended to support public health and well-being. Keeping patients at the centre of decision-making, especially governance, can improve trust and ensure that AI-guided technologies meet the needs and address the concerns of patients.

6. Addressing challenges to ethics and governance

Several intergovernmental and United Nations agencies have proposed standards and recommendations to improve the governance of AI in health.

UNESCO, in its report on Big Data and Health, identified six elements to be considered to ensure an appropriate governance framework for AI: procedures for consent (and alternative models) and dissent; transparency of algorithms; disclosure of commercial interests and collaboration with commercial parties; benefit-sharing; provisions for indigenous and local populations and traditional minorities; and provisions for children and adolescents (informed consent at the age of maturity).

Frameworks for the production and management of data have also been proposed. In 2016, the OECD issued recommendations for health data governance. These recommendations, which can be applied to AI, are intended to address three major policy challenges:

1. balancing availability and use of personal health data for health-related public interest purposes and promoting the protection of privacy, personal health data and data security;
2. reinforcing trust and empowering users through proactive stakeholder engagement and community building; and
3. encouraging multi-country statistical and research projects and research uses of data for the public interest and protecting privacy and data security.

Panellists discussed a variety of mechanisms and practices that could be used to encourage ethical development and use of AI in various practices that are emerging in the use of AI in health care.

Role of design in ensuring appropriate development and use of artificial intelligence

Ethics can and should be integrated into the design of a technology as ethics and engineering come ever closer. To ensure that engineering and design address a wide range of ethical concerns, the concepts to be supported should be identified, such as privacy and security, and the ethical terms should be defined as specifically as possible for use by scientists and engineers in developing a technology. Design also includes making a technology work and a system of governance (rules and norms) applicable to the context in which the technology is used.

Focusing on the ethical dimensions of design is an opportunity to consider technology not only as part of the problem but also as part of the solution, as it can be used to address numerous ethical concerns. Design must be grounded in values that can be translated into specific design requirements, and it must ensure that specific responsibility and obligations are created to uphold ethical values. Taken together, these elements can promote responsible innovation,
wherein a technology can be designed without compromising ethics, by pursuing moral goals and combining the goals into one design.

**Box 5. Responsible innovation**

1. Identify serious problems to be solved.
2. Think in advance about the consequences of and alternatives to proposed solutions.
3. Evaluate solutions and alternatives in terms of relevant moral values (including cost, cost-saving and efficiency).
4. Secure direct and indirect help from a broad range of stakeholders.
5. Use the resulting moral considerations as design requirements

**Importance of social and legal systems and context to ensure ethical use of artificial intelligence**

The design of a technology, such as an algorithm and the wider social and legal contexts, ethical concerns and underlying causes of harm must be considered in the development, design and introduction of technologies. Thus, in ethical design and implementation of AI in health systems, attention must be paid to the many questions that such systems raise for justice in health care and for health system governance and accountability.

One participant noted that human rights are not protected by technology but by a set of legal and social norms and rules. Trying to engineer “fairness” and justice into machine learning algorithms and models by using fairness itself as a property of the system is a category error. Fairness and justice are not properties of the technical tools themselves but can be addressed only through the social and legal systems in which technologies are used.

Thus, focus on the technology itself and on some of its impacts, such as bias, may result in losing sight of the broader systemic issues that may be the main cause of harm. Such focus can also divert attention from the roles and responsibilities of institutions and entities that have tremendous power, including companies, governments and policy-making bodies.

Designers can address the shortcomings of technology and avoid traps that raise ethical concern or prevent broader uptake of a technology that provides a significant health benefit by focusing on design as a process rather than on the solutions and by ensuring that the design integrates social actors and social processes and not just technical challenges. For example, better communication between clinicians and data scientists could improve quality and sustainability. The following sections address the social actors and social processes that can be used to promote ethical AI and to address the underlying harms and direct impacts of technology on individuals, health care systems and societies.

**The centrality of human rights**

Although the Universal Declaration of Human Rights was written over 70 years ago, the principles, the Declaration and subsequent human rights law and standards all provide solid foundations for ethical frameworks for AI and Big Data in health care, including provisions for accountability and remedy.

The Council of Europe has built its guidance on AI and AI in health care mainly on the basic norms and obligations in international, regional and national human rights standards. Similarly, the Toronto Declaration, convened by major human rights and digital rights organizations, including Amnesty International and Access Now, calls upon governments, foundations and
companies to uphold a wide range of obligations and responsibilities under human rights laws and standards and to avoid discrimination in how AI is used. These standards, as one panellist noted, are entirely applicable and relevant to the use of AI in health care.

Human rights can provide substantive standards to guide ethical frameworks, laws and policies and the design of technology and can also be used to assess the positive and negative impacts of AI technologies in risk assessments, for example. Furthermore, human rights laws and standards can set standards and norms for transparency and accountability, including how automated and machine learning decisions are reached. A set of principles, the United Nations Guiding Principles for Business and Human Rights, lists the core obligations of business to protect, respect and fulfil human rights.

Importance of public debate and dialogue in earning trust

Several panellists noted that public engagement and dialogue ensure that AI meets certain core societal expectations for its use in health care. Furthermore, effective public engagement could engender greater trust and acceptance of AI. Public dialogue ensures ascertainment of society’s views, as far as possible, on the ethical dimensions of AI, its design and uses.

There have been many examples of public debate and dialogue on the topic in the United Kingdom of Great Britain and Northern Ireland. The aim of Health Data Research, an independent, not-for-profit organization of 22 research institutions across the United Kingdom of Great Britain and Northern Ireland, is to collect health data throughout the country and to make it available to public and private entities to understand diseases and find ways to prevent, treat and cure them. It has found that public engagement is necessary and critical for building trust. The organization is working with partners, including the Wellcome Trust initiative Understanding Patient Data (50). Workshops have provided a forum for citizens to discuss their expectations and concerns with use of patient data in AI and other applications. Before these workshops, sponsored by the Health Research Authority, 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 increased to 45% (51). The survey also found that patients viewed contributing data as a value exchange, with a societal benefit, and wanted the National Health Service to benefit from their data. They further considered it acceptable for commercial companies to have access to their data, provided that the benefit returned to the public, and that the National Health Service administer the data for the public benefit.

Similarly, a meeting co-hosted by the United Kingdom Academy of Medical Sciences found 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” (52). The Academy has conducted public dialogue on the “data driven future” in order to understand awareness, expectations, aspirations and concerns about future technologies that require that patient data be accessed, analysed or linked for clinical diagnosis and management. 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 National Health Service) should oversee the data and administer it for the public benefit (53).

Building trust: human warranty, testing and validation

The panellists agreed that if transparency and explainability are not only increasingly difficult to achieve as algorithms become more complex and more difficult to scrutinize, and if
explainability is not always desirable (see above), there must be other mechanisms to build the societal trust that is required to ensure that AI is acceptable and trustworthy and that it is designed and implemented in a manner that conforms to societal expectations.

Careful empirical validation of the practical merits of an intervention can build trust, ensure that resources are well spent and contribute to medical advances. Regulatory practices should ensure that AI systems are used only for the tasks for which they or their accuracy have been validated. The cases of use for which a machine learning system is suited and validated must therefore be designated and uses for which it has not been validated should be discouraged. Each technology must be tested in the diverse real-world contexts in which it is to be applied, and its performance assessed against that of humans in the same decision task in well-designed empirical studies.

Even after a technology is deployed, it should undergo “continuous quality improvement”, whereby its performance is audited and assessed according to changes in the external environment. In France, for example, trust in AI is built by focusing on human warranty, to avoid the difficulties associated with explainability. Human warranty requires evaluation by patients and clinicians at critical points in the development and deployment of AI. Human warranty can involve periodic meetings of practitioners, health care providers and patient representatives to collect data, de-identify and evaluate false-positive and false-negative results obtained with the data and determine whether the algorithm is adequate. The human warranty principle has been integrated into Article 11 of the French Bioethics Law. Panellists endorsed the principle as embodying the very idea of AI for positive health care regulation.

In LMIC, the “solution trap” must be avoided to build trust and ensure adoption and use of AI. Deployment of AI should be guided by a problem to be solved and the AI should be best suited to solving it, as in high-income countries. The right skills must be available for appropriate use of AI in these contexts, including digital skills, in the health workforce and updated curricula for practitioners and technology designers. Appropriate investments in medical education will be critical to ensure that AI is used appropriately. More “citizen science” and citizen involvement is necessary to encourage appropriate uses and adoption of AI in health care.

Preserving the role of the State in the age of artificial intelligence and Big Data

There is concern that AI will undermine the ability of individuals and practitioners to make appropriate medical decisions and that the broader decision-making authority of medical systems and governments will be increasingly surrendered to the companies that develop and deploy AI for health care or to the technology itself (e.g. governance by data).

Algorithmic governance, in the broadest sense, would avoid any confrontation or engagement with individuals, whether through debates, consultation or elections, including how health care is developed and deployed and how resources are used. A surrender of control is possible because of people’s willingness to delegate a range of decisions to AI, with less priority given to democratic debate than to real-time processing of data through sophisticated algorithms.

When most of the data and algorithms are managed by large technology companies, it will be increasingly likely that those companies will govern decisions that should be in the domain of citizens, societies and governments, because of their control and power over the resources and information that underpin the digital economy. Thus, one panellist proposed that the questions to be answered are: “What is the role of the State in the AI political economy?” and “What should governments do to rebalance power from the private sector to individual users?”
These questions will be considered by the expert group at subsequent meetings. They are challenges that must also be taken up elsewhere to ensure appropriate use of AI in health care and appropriate governance, ownership and authority over its use.

7. Consensus points

The WHO guidance on ethics and governance of AI for health will be prepared in order to harness AI ethically to improve the health and well-being of individuals, to save lives and, more broadly, to serve global health. The guidance must be grounded in human rights and other established instruments such as ethical guidelines for research, building on existing guidance and based on ethical principles that underpin governance. Ethical principles and frameworks in related disciplines, such as computer science, must also be considered.

The guidance will be practical, with use cases and country experience, and will be applicable in various health care systems (including in LMIC) and associated with existing health care strategies. The guidance will be “people-centred” and will be as useful for regulators as it is for health care professionals. While WHO guidance is primarily intended for ministries of health, it must reach other key government ministries (such as information technology), the private sector and the public (including patient and civil society organizations) if it is to have a true impact.

Participants stressed the importance of the guidance for shaping the ethical questions of research agendas on digital health and AI in scientific academies and national institutes.

The guidance is intended to provide ethical guidance in the short, medium and long term and to be open to frequent revision. It was noted that there are risks associated with excessive specificity in any guidance document, which could stifle beneficial innovation. A framework for guidance should be structured either according to stages of clinical care and medicine – such as prevention, care, treatment and monitoring – or on the various categories of public health practice used by governments and health care systems. In general, there was agreement that the principles must be easy to adopt into existing government guidance for operating health care systems. Other speakers reminded the group that, even though the recommendations of the framework are intended to be built on the uses of AI in health, the technologies themselves are composed of inputs generated by programmers and coders that are based on the collection and use of data. The rules of engagement with the private sector and the role of the private sector within AI for health were themes that emerged repeatedly throughout the consultation and should be addressed in the guidance document.

Participants listed several principles that must be included in the guidance, which are transparency, trustworthiness, privacy, fairness, human warranty and sustainability. Other topics that could be included, such as data quality, are likely to be addressed by a technical advisory group of the WHO Department of Digital Health. The two groups should cross-reference each other’s work and recommendations.

8. Conclusion

The meeting brought together a multidisciplinary group of ethicists, doctors, scientists, policymakers, lawyers, civil society representatives, human rights experts and industry. They heard experts’ overviews of the use cases of AI in health and the ethical dimensions and challenges to be addressed if AI is scaled up for use around the world in a fair, just, equitable manner.

As noted above, similar yet separate processes are providing governments and non-State actors with concrete guidance to ensure ethical use of AI technologies. WHO welcomes and is
contributing to this work and hopes to open additional channels of collaboration and discussion
to identify possible common recommendations.

WHO’s future guidance will play a critical role in international work to ensure that AI makes a
positive contribution to global health and is successfully integrated into health care systems at
all levels of development. Subsequent meetings will address other ethical and governance
challenges associated with use of AI in health care and begin to formulate concrete policy
guidance to address the many risks and benefits of AI in health care systems, to health care
practitioners and, most importantly, to the millions of people who seek affordable, appropriate,
equitable health care worldwide.

Input received at this meeting provides a foundation for WHO, which will convene several
follow-up meetings in 2020, to prepare a guidance document that can be used by a wide range
of stakeholders in the development and use of AI in health care. WHO anticipates that the
guidance document will be completed in 2020.
References


17. AINDRA. Bangalore; 2020 (www.ain dra.in, accessed February 2020).


21. The web application was an example mentioned during the expert consultation.


Annex 1. List of participants

Experts

Ms Maria Paz Canales, Executive Director, Digitales PGP Fingerprint, Santiago, Chile

Dr Arisa Ema, Ito International Research Center, Institute for Future Initiative, Tokyo, Japan (remotely)

Dr Amel Ghouila, Bioinformatician, Capacity Builder at H3ABioNet, Tunis, Tunisia

Professor Kenneth Goodman, Institute for Bioethics and Health Policy, University of Miami Miller, Miami (FL), United States of America

Professor Malavika Jayaram, Executive Director, Digital Asia Hub, Hong Kong Special Administrative region, China (remotely)

Professor Daudi Jjingo, Scientist and Lecturer, Makerere University, College of Computing, H3Africa Bioinformations Education Grant, Director, African Centre of Excellence in Bioinformatics and Data Science, Kampala, Uganda (remotely)

Professor Tze Yun Leong, School of Computing, National University of Singapore, Singapore

Professor Alex John London, Philosophy Department, Carnegie Mellon University, Pittsburgh (PA), United States of America (remotely)

Professor Partha Majumder, Distinguished Professor, National Institute of Biomedical Genomics, Kalyani, India

Professor Tshilidzi Marwala, University of Johannesburg, Johannesburg, South Africa (remotely)

Professor Roli Mathur, Scientist and Head, ICMR Bioethics Unit, National Centre for Disease Informatics and Research, Bangalore, India

Professor Andrew Morris, Health Data Research United Kingdom, London, United Kingdom of Great Britain and Northern Ireland

Professor Hamid R. Rabiee, Sharif University of Technology, Department of Computer Engineering, Tehran, Islamic Republic of Iran

Professor M. Jeroen Van den Hoven, Delft University of Technology, Delft, The Netherlands

Professor Effy Vayena, Chair of Bioethics, Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

Professor Robyn Whittaker, Waitemata District Health Board and National Institute for Health Innovation, University of Auckland, Auckland, New Zealand

Professor Yi Zeng, Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Observers

Professor Abdoulaye Baniré Diallo, Department of Computer Science, Université du Québec à Montréal Co-Lead, Acces Omic Senegal, Montreal, Quebec, Canada
Mrs Elizabeth Bohm, United Kingdom Academy of Medical Sciences, London, United Kingdom of Great Britain and Northern Ireland

Mr Julien Durand, IFPMA, Executive Director, Compliance, Amgen Switzerland AG, Switzerland

Dr Osama El-Hassan, Rashid Hospital, Dubai Health Authority, Dubai, United Arab Emirates

Professor David Gruson, Executive Committee, Health Sciences, Paris, France

Mr Lee Hibbard, Bioethics Unit, Council of Europe, Strasbourg, France

Dr Lauren Milner, Center for Drug Evaluation and Research, Office of Medical Policy, Division of Clinical Trial Quality, Food and Drug Administration, Silver Spring (MD), United States of America (remotely)

Dr M. Khair El Zarrad, Center for Drug Evaluation and Research, Office of Medical Policy, Division of Clinical Trial Quality, Food and Drug Administration, Silver Spring (MD), United States of America (remotely)

Ms Daniela Paolotti, ISI Foundation, Turin, Italy

Dr Rasha Abdul Rahim, Deputy Director, Amnesty International, Technology and Human Rights, London, United Kingdom of Great Britain and Northern Ireland

Dr Elettra Ronchi, OECD, Paris, France

Mr Tee Wee Ang, UNESCO, Paris, France

**WHO headquarters**

Ms Kidist K. Bartolomeos DGO/SPI/WHP
Dr Marie-Charlotte Bouësseau HIS/SDS
Dr Somnath Chatterji HMM/MAA
Dr Philippa Easterbrook CDS/HIV/GHP
Ms Marisol Guraiib HIS/IER/REK
Dr Garrett Livingston Mehl FWC/RHR/AGH
Ms Katherine Littler HIS/IER/REK
Ms Lee-Anne Pascoe, WHO intern
Mr Amit Prasad HIS/IER
Dr Sameer Pujari NMH/PND
Dr John Reeder CDS/TDR/DIR
Dr Andreas Reis HIS/IER/REK
Dr Vasee Moorthy HIS/IER/REK
Ms Diana Zandi HIS/SDS
Mr Rohit Malpani, Consultant
Ms Tracy Swan, Rapporteur
Geneva, Switzerland

**WHO regional offices**

Dr Ahmed Mandil, Coordinator, Research Development and Innovation, WHO Regional Office for the Eastern Mediterranean (remotely)
Mr Mohamed Hassan Nour, National Professional Officer, WHO Regional Office for the Eastern Mediterranean (remotely)
Mr Clayton Hamilton, Regional adviser, Digital Health, WHO Regional Office for Europe

All invited experts completed the Declaration of interests form for WHO experts before the meeting. The Health Ethics and Governance Unit, in consultation with the WHO Office of Compliance, Risk
Management and Ethics, assessed the interests declared by the experts. Most experts declared no interests, and others declared interests that were considered as minimal or insignificant.
Annex 2. Agenda

Day 1. Wednesday, 2 October 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00–</td>
<td>Welcoming address</td>
<td>Clayton Hamilton, on behalf of Bernardo Mariano, Director, WHO Digital</td>
</tr>
<tr>
<td>09:30</td>
<td>Introductions</td>
<td>Health Department</td>
</tr>
<tr>
<td>09:30–</td>
<td>Declarations of Interest</td>
<td>Andreas Reis &amp; Rohit Malpani</td>
</tr>
<tr>
<td>09:45</td>
<td>Goals of the project and outline of the first meeting</td>
<td></td>
</tr>
<tr>
<td>10:30–</td>
<td>Overview presentation and group discussion about any gaps as to</td>
<td>Effy Vayena</td>
</tr>
<tr>
<td>10:00</td>
<td>salient issues/challenges related to the use of AI in health.</td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:00–</td>
<td>Mobile health</td>
<td>Clayton Hamilton</td>
</tr>
<tr>
<td>11:15</td>
<td>How AI will impact drug development</td>
<td>Julien Durand</td>
</tr>
<tr>
<td>11:30</td>
<td>AI and public health surveillance</td>
<td>Daniela Paolotti</td>
</tr>
<tr>
<td>11:45</td>
<td>From diagnosis to prediction</td>
<td>Partha Majumder</td>
</tr>
<tr>
<td>12:00</td>
<td>Group discussion</td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Additional group discussion</td>
<td></td>
</tr>
<tr>
<td>14:30</td>
<td>Horizon scanning as to how AI for health will evolve.</td>
<td>Leong Tze Yun</td>
</tr>
<tr>
<td>15:00</td>
<td>Group discussion</td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Panel and group discussion regarding guidance issued by institutions</td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>and inter-governmental organizations.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Panelists:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tee Wee Ang, UNESCO</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeroen van den Hoeven (on behalf of the European Commission)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lee Hibbard, Council of Europe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elettra Ronchi, OECD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shakeel Bhatti, WIPO</td>
<td></td>
</tr>
<tr>
<td></td>
<td>End of day 1</td>
<td></td>
</tr>
</tbody>
</table>
## Day 2. Thursday, 3 October 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00–09:30</td>
<td>General overview with a focus on bias, discrimination and fairness</td>
<td>Jeroen van den Hoeven</td>
</tr>
<tr>
<td>09:30–10:00</td>
<td>The digital divide and equitable access</td>
<td>Amel Ghouila</td>
</tr>
<tr>
<td>10:00–10:30</td>
<td>Reconciling AI with human rights and dignity in low- and middle-income countries</td>
<td>Malavica Jayaram</td>
</tr>
<tr>
<td>10:30–11:00</td>
<td>The physician/machine – patient relationship and the transformation of clinical care</td>
<td>Robyn Whittaker</td>
</tr>
<tr>
<td>11:00–11:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:30–12:00</td>
<td>Data confidentiality, privacy and informed consent</td>
<td>Ken Goodman</td>
</tr>
<tr>
<td>12:00–12:30</td>
<td>Group discussion</td>
<td></td>
</tr>
<tr>
<td>12:30–13:30</td>
<td><strong>Lunch</strong></td>
<td></td>
</tr>
<tr>
<td>13:30–14:00</td>
<td>Algorithms: Explainability and transparency</td>
<td>Alex John London</td>
</tr>
<tr>
<td>14:00–14:30</td>
<td>Data sharing</td>
<td>Andrew Morris</td>
</tr>
<tr>
<td>14:30–15:00</td>
<td>Data analysis</td>
<td>Partha Majumder</td>
</tr>
<tr>
<td>15:00–15:30</td>
<td>Review of relevant principles related to AI: Identifying consensus principles for AI</td>
<td>Effy Vayena</td>
</tr>
<tr>
<td>15:30–15:45</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>15:45–17:45</td>
<td>Group discussion on which general guiding principles for AI relate to health care and how such principles apply particularly within the health care field</td>
<td></td>
</tr>
<tr>
<td>17:45</td>
<td>End of day 2</td>
<td></td>
</tr>
</tbody>
</table>
## Day 3. Friday, 4 October 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30–10:30</td>
<td>Panel and group discussion on building trustworthy AI: Engaging the public, health care professionals and users Panellists: Elizabeth Bohm, United Kingdom Academy of Medical Sciences Rasha Abdul Rahim, Amnesty International David Gruson, Ethik IA</td>
<td></td>
</tr>
<tr>
<td>10:30–10:45</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>10:45–11:45</td>
<td>Agree on main principles for ethical framework and the outline and approach of a guidance document</td>
<td></td>
</tr>
<tr>
<td>11:45–12:45</td>
<td>Determine priorities for the subsequent meetings.</td>
<td></td>
</tr>
<tr>
<td>12:45–13:00</td>
<td>Closing remarks</td>
<td>John Reeder, Director, WHO Research for Health Department</td>
</tr>
<tr>
<td>13:00</td>
<td>End of consultation</td>
<td></td>
</tr>
</tbody>
</table>