1. INTRODUCTION

Digital technologies and artificial intelligence (AI), particularly machine learning, are transforming medicine, medical research and public health. Technologies based on AI are now used in health services in countries of the Organization for Economic Co-operation and Development (OECD), and its utility is being assessed in low- and middle-income countries (LMIC). The United Nations Secretary-General has stated that safe deployment of new technologies, including AI, can help the world to achieve the United Nations Sustainable Development Goals (*1*), which would include the health-related objectives under Sustainable Development Goal 3. AI could also help to meet global commitments to achieve universal health coverage.

Use of AI for health nevertheless raises trans-national ethical, legal, commercial and social concerns. Many of these concerns are not unique to AI. The use of software and computing in health care has challenged developers, governments and providers for half a century, and AI poses additional, novel ethical challenges that extend beyond the purview of traditional regulators and participants in health-care systems. These ethical challenges must be adequately addressed if AI is to be widely used to improve human health, to preserve human autonomy and to ensure equitable access to such technologies.

Use of AI technologies for health holds great promise and has already contributed to important advances in fields such as drug discovery, genomics, radiology, pathology and prevention. AI could assist health-care providers in avoiding errors and allow clinicians to focus on providing care and solving complex cases. The potential benefits of these technologies and the economic and commercial potential of AI for health care presage ever greater use of AI worldwide.

Unchecked optimism in the potential benefits of AI could, however, veer towards habitual first recourse to technological solutions to complex problems. Such "techno-optimism" could make matters worse, for example, by exacerbating the unequal distribution of access to health-care technologies within and among wealthy and low-income countries (*2*). Furthermore, the digital divide could exacerbate inequitable access to health-care technologies by geography, gender, age or availability of devices, if countries do not take appropriate measures. Inappropriate use of AI could also perpetuate or exacerbate bias. Use of limited, low-quality, non-representative data in AI could perpetuate and deepen prejudices and disparities in health care. Biased inferences, misleading data analyses and poorly designed health applications and tools could be harmful. Predictive algorithms based on inadequate or inappropriate data can result in significant racial or ethnic bias. Use of high-quality, comprehensive datasets is essential.

AI could present a singular opportunity to augment and improve the capabilities of over-stretched health-care workers and providers. Yet, the introduction of AI for health care, as in many other sectors of the global economy, could have a significant negative impact on the health-care workforce. It could reduce the size of the workforce, limit, challenge or degrade the skills of health workers, and oblige them to retrain to adapt to the use of AI. Centuries of medical practice are based on relationships between provider and patient, and particular care must be taken when introducing AI technologies so that they do not disrupt such relationships.

The Universal Declaration of Human Rights, which includes pillars of patient rights such as dignity, privacy, confidentiality and informed consent, might be dramatically redefined or undermined as digital technologies take hold and expand. The performance of AI depends (among other factors) on the nature, type and volume of data and associated information and the conditions under which such data were gathered. The pursuit of data, whether by government or companies, could undermine privacy and autonomy at the service of government or private surveillance or commercial profit. If privacy and autonomy are not assured, the resulting limitation of the ability to exercise the full range of human rights, including civil and political rights (such as freedom of movement and expression) and social and economic rights (such as access to health care and education), might have a wider impact.

AI technologies, like many information technologies used in health care, are usually designed by companies or through public–private partnerships (PPPs), although many governments also develop and deploy these technologies. Some of the world's largest technology companies are developing new applications and services, which they either own or invest in. Many of these companies have already accumulated large quantities of data, including health data, and exercise significant power in society and the economy. While these companies may offer innovative approaches, there is concern that they might eventually exercise too much power in relation to governments, providers and patients.

AI technologies are also changing where people access health care. AI technologies for health are increasingly distributed outside regulated health-care settings, including at the workplace, on social media and in the education system. With the rapid proliferation and evolving uses of AI for health care, including in response to the COVID-19 pandemic, government agencies, academic institutions, foundations, nongovernmental organizations and national ethics committees are defining how governments and other entities should use and regulate such technologies effectively. Ethically optimized tools and applications could sustain widespread use of AI to improve human health and the quality of life, while mitigating or eliminating many risks and bad practices.

To date, there is no comprehensive international guidance on use of AI for health in accordance with ethical norms and human rights standards. Most countries do not have

laws or regulations to regulate use of AI technologies for health care, and their existing laws may not be adequate or specific enough for this purpose. WHO recognizes that ethics guidance based on the shared perspectives of the different entities that develop, use or oversee such technologies is critical to build trust in these technologies, to guard against negative or erosive effects and to avoid the proliferation of contradictory guidelines. Harmonized ethics guidance is therefore essential for the design and implementation of AI for global health.

The primary readership of this guidance document is ministries of health, as it is they that determine how to introduce, integrate and harness these technologies for the public good while restricting or prohibiting inappropriate use. The development, adoption and use of AI nevertheless requires an integrated, coordinated approach among government ministries beyond that for health. The stakeholders also include regulatory agencies, which must validate and define whether, when and how such technologies are to be used, ministries of education that teach current and future health-care workforces how such technologies function and are to be integrated into everyday practice, ministries of information technology that should facilitate the appropriate collection and use of health data and narrow the digital divide and countries' legal systems that should ensure that people harmed by AI technologies can seek redress.

This guidance document is also intended for the stakeholders throughout the health-care system who will have to adapt to and adopt these technologies, including medical researchers, scientists, health-care workers and, especially, patients. Access to such technologies can empower people who fall ill but can also leave them vulnerable, with fewer services and less protection. People have always been at the centre at all levels of decision-making in health care, whereas the inevitable growth of AI for health care could eventually challenge human primacy over medicine and health.

This guidance is also designed for those responsible for the design, deployment and refinement of AI technologies, including technologists and software developers. Finally, it is intended to guide the companies, universities, medical associations and international organizations that will, with governments and ministries of health, set policies and practices to define use of AI in the health sector. In identifying the many ethical concerns raised by AI and by providing the relevant ethical frameworks to address such concerns, this document is intended to support responsible use of AI worldwide.

WHO recognizes that AI is a fast-moving, evolving field and that many applications, not yet envisaged, will emerge as ever-greater public and private investment is dedicated to the use of AI for health. For example, in 2020, WHO issued interim guidance on the use of proximity tracking applications intended to facilitate contact-tracing during the COVID-19 pandemic. WHO may consider specific guidance for additional tools and applications and periodically update this guidance to keep pace with this rapidly changing field.

2. ARTIFICIAL INTELLIGENCE

"Artificial intelligence" generally refers to the performance by computer programs of tasks that are commonly associated with intelligent beings. The basis of AI is algorithms, which are translated into computer code that carries instructions for rapid analysis and transformation of data into conclusions, information or other outputs. Enormous quantities of data and the capacity to analyse such data rapidly fuel AI (*3*). A specific definition of AI in a recommendation of the Council on Artificial Intelligence of the OECD (*4*) states:

> An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The various types of AI technology include machine-learning applications such as pattern recognition, natural language processing, signal processing and expert systems. Machine learning, which is a subset of AI techniques, is based on use of statistical and mathematical modelling techniques to define and analyse data. Such learned patterns are then applied to perform or guide certain tasks and make predictions.

Machine learning can be subcategorized according to how it learns from data into supervised learning, unsupervised learning and reinforced learning. In supervised learning, data used to train the model are labelled (the outcome variable is known), and the model infers a function from the data that can be used for predicting outputs from different inputs. Unsupervised learning does not involve labelling data but involves identification of hidden patterns in the data by a machine. Reinforcement learning involves machine learning by trial and error to achieve an objective for which the machine is "rewarded" or "penalized", depending on whether its inferences reach or hinder achievement of an objective (*5*). Deep learning, also known as "deep structured learning", is a family of machine learning based on use of multi-layered models to progressively extract features from data. Deep learning can be supervised, unsupervised or semi-supervised. Deep learning generally requires large amounts of data to be fed into the model.

Many machine-learning approaches are data-driven. They depend on large amounts of accurate data, referred to as "big data", to produce tangible results. "Big data" are complex data that are rapidly collected in such unprecedented quantities that terabytes (one trillion units [bytes] of digital information), petabytes (1000 terabytes)

or even zettabytes (one million petabytes) of storage space may be required as well as unconventional methods for their handling. The unique properties of big data are defined by four dimensions: volume, velocity, veracity and variety.

AI could improve the delivery of health care, such as prevention, diagnosis and treatment of disease (*6*), and is already changing how health services are delivered in several high-income countries (HIC). The possible applications of AI for health and medicine are expanding continually, although the use of AI may be limited outside HIC because of inadequate infrastructure. The applications can be defined according to the specific goals of use of AI and how AI is used to achieve those goals (methods). In health care, usable data have proliferated as a result of collection from numerous sources, including wearable technologies, genetic information generated by genome sequencing, electronic health-care records, radiological images and even from hospital rooms (*7*).

3. APPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR HEALTH

This section identifies AI technologies developed and used in HIC, although examples of such technologies are emerging (and being pilot-tested or used) in LMIC. Digital health technologies are already used widely in LMIC, including for data collection, dissemination of health information by mobile phones and extended use of electronic medical records on open-software platforms and cloud computing (*8*). Schwabe and Wahl (*9*) have identified four uses of AI for health in LMIC: diagnosis, morbidity or mortality risk assessment, disease outbreaks and surveillance, and health policy and planning.

3.1 In health care

The use of AI in medicine raises notions of AI replacing clinicians and human decision-making. The prevailing sentiment is, however, that AI is increasingly improving diagnosis and clinical care, based on earlier definitions of the role of computers in medicine (*10*) and regulations in which AI is defined as a support tool (to improve judgement).

Diagnosis and prediction-based diagnosis

AI is being considered to support diagnosis in several ways, including in radiology and medical imaging. Such applications, while more widely used than other AI applications, are still relatively novel, and AI is not yet used routinely in clinical decision-making. Currently, AI is being evaluated for use in radiological diagnosis in oncology (thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging and dose optimization for radiological treatment), in non-radiological applications (dermatology, pathology), in diagnosis of diabetic retinopathy, in ophthalmology and for RNA and DNA sequencing to guide immunotherapy (*11*). In LMIC, AI may be used to improve detection of tuberculosis in a support system for interpreting staining images (*12*) or for scanning X-rays for signs of tuberculosis, COVID-19 or 27 other conditions (*13*).

Nevertheless, few such systems have been evaluated in prospective clinical trials. A recent comparison of deep-learning algorithms with health-care professionals in detection of diseases by medical imaging showed that AI is equivalent to human medical judgement in specific domains and applications in specific contexts but also that "few studies present externally validated results or compare the performance of deep learning models and health-care professionals using the same sample" (*14*). Other questions are whether the performance of AI can be generalized to implementation in practice and whether AI trained for use in one context can be used accurately and safely in a different geographical region or context.

As AI improves, it could allow medical providers to make faster, more accurate diagnoses. AI could be used for prompt detection of conditions such as stroke, pneumonia, breast cancer by imaging (*15*, *16*), coronary heart disease by echocardiography (*17*) and detection of cervical cancer (*18*). Unitaid, a United Nations agency for improving diagnosis and treatment of infectious diseases in LMIC, launched a partnership with the Clinton Health Access Initiative in 2018 to pilot-test use of an AI-based tool to screen for cervical cancer in India, Kenya, Malawi, Rwanda, South Africa and Zambia (*19*). Many low-income settings facing chronic shortages of health-care workers require assistance in diagnosis and assessment and to reduce their workload. It has been suggested that AI could fill gaps in the absence of health-care services or skilled workers *(9*).

AI might be used to predict illness or major health events before they occur. For example, an AI technology could be adapted to assess the relative risk of disease, which could be used for prevention of lifestyle diseases such as cardiovascular disease (*20*, *21*) and diabetes (*22*). Another use of AI for prediction could be to identify individuals with tuberculosis in LMIC who are not reached by the health system and therefore do not know their status (*23*). Predictive analytics could avert other causes of unnecessary morbidity and mortality in LMIC, such as birth asphyxia. An expert system used in LMIC is 77% sensitive and 95% specific for predicting the need for resuscitation (*8*). Several ethical challenges to prediction-based health care are discussed in section 6.5.

Clinical care

Clinicians might use AI to integrate patient records during consultations, identify patients at risk and vulnerable groups, as an aid in difficult treatment decisions and to catch clinical errors. In LMIC, for example, AI could be used in the management of antiretroviral therapy by predicting resistance to HIV drugs and disease progression, to help physicians optimize therapy (23). Yet, clinical experience and knowledge about patients is essential, and AI will not be a substitute for clinical due diligence for the foreseeable future. If it did, clinicians might engage in "automation bias" and not consider whether an AI technology meets their needs or those of the patient. (See section 6.4.)

The wider use of AI in medicine also has technological challenges. Although many prototypes developed in both the public and the private sectors have performed well in field tests, they often cannot be translated, commercialized or deployed. An additional obstacle is constant changes in computing and information technology management, whereby systems become obsolete ("software erosion") and companies disappear. In resource-poor countries, the lack of digital infrastructure and the digital divide (See section 6.2.) will limit use of such technologies.

Health-care workers will have to adapt their clinical practice significantly as use of AI increases. AI could automate tasks, giving doctors time to listen to patients, address their fears and concerns and ask about unrelated social factors, although they may still worry about their responsibility and accountability. Doctors will have to update their competence to communicate risks, make predictions and discuss trade-offs with patients and also express their ethical and legal concern about understanding AI technology. Even if technology makes the predicted gains, those gains will materialize only if the individuals who manage health systems use them to extend the capacity of the health system in other areas, such as better availability of medicines or other prescribed interventions or forms of clinical care.

Emerging trends in the use of AI in clinical care

Several important changes imposed by the use of AI in clinical care extend beyond the provider–patient relationship. Four trends described here are: the evolving role of the patient in clinical care; the shift from hospital to home-based care; use of AI to provide "clinical" care outside the formal health system; and use of AI for resource allocation and prioritization. Each of these trends has ethical implications, as discussed below.

The evolving role of the patient in clinical care

AI could eventually change how patients self-manage their own medical conditions, especially chronic diseases such as cardiovascular diseases, diabetes and mental problems (*24*). Patients already take significant responsibility for their own care, including taking medicines, improving their nutrition and diet, engaging in physical activity, caring for wounds or delivering injections. AI could assist in self-care, including through conversation agents (e.g. "chat bots"), health monitoring and risk prediction tools and technologies designed specifically for individuals with disabilities (24). While a shift to patient-based care may be considered empowering and beneficial for some patients, others might find the additional responsibility stressful, and it might limit an individual's access to formal health-care services.

The growing use of digital self-management applications and technologies also raises wider questions about whether such technologies should be regulated as clinical applications, thus requiring greater regulatory scrutiny, or as "wellness applications", requiring less regulatory scrutiny. Many digital self-management technologies arguably fall into a "grey zone" between these two categories and may present a risk if they are used by patients for their own disease management or clinical care but remain largely unregulated or could be used without prior medical advice. Such concerns are exacerbated by the distribution of such applications by entities that are not a part of the formal health-care system. This related but separate trend is discussed below.

The shift from hospital to home-based care

Telemedicine is part of a larger shift from hospital- to home-based care, with use of AI technologies to facilitate the shift. They include remote monitoring systems, such as video-observed therapy for tuberculosis and virtual assistants to support patient care. Even before the COVID-19 pandemic, over 50 health-care systems in the USA were making use of telemedicine services (*25*). COVID-19, having discouraged people in many settings from visiting health-care facilities, accelerated and expanded the use of telemedicine in 2020, and the trend is expected to continue. In China, the number of telemedicine providers has increased by nearly four times during the pandemic (*26*).

The shift to home-based care has also partly been facilitated by increased use of search engines (which rely on algorithms) for medical information as well as by the growth in the number of text or speech chatbots for health care (*27*), the performance of which has improved with improvements in natural language processing, a form of AI that enables machines to understand human language. The use of chatbots has also accelerated during the COVID-19 pandemic (*28*).

Furthermore, AI technologies may play a more active role in the management of patients' health outside clinical settings, such as in "just-in-time adaptive interventions". These rely on sensors to provide patients with specific interventions according to data collected previously and currently; they also notify a health-care provider of any emerging concern (*29*). The growth and use of sensors and wearables may improve the effectiveness of "just-in-time adaptive interventions" but also raise concern, in view of the amount of data such technologies are collecting, how they are used and the burden such technologies may shift to patients.

Use of AI to extend "clinical" care beyond the formal health-care system

AI applications in health are no longer exclusively used in health-care systems (or home care), as AI technologies for health can be readily acquired and used by non-health system entities. This has meant that people can now obtain health-care services outside the health-care system. For example, AI applications for mental health are often provided through the education system, workplaces and social media and may even be linked to financial services (*30*). While there may be support for such extended uses of health applications to compensate for both increased demand and a limited number of providers (31), they generate new questions and concerns. (See section 9.3.)

These three trends may require near-continuous monitoring (and self-monitoring) of people, even when they are not sick (or are "patients"). AI-guided technologies require the use of mobile health applications and wearables, and their use has increased with the trend to self-management (*31*). Wearable technologies include those placed in the body (artificial limbs, smart implants), on the body (insulin pump patches, electroencephalogram devices) or near the body (activity trackers, smart watches and

smart glasses). By 2025, 1.5 billion wearable units may be purchased annually. [1] Wearables will create more opportunities to monitor a person's health and to capture more data to predict health risks, often with greater efficiency and in a timelier manner.

Although such monitoring of "healthy" individuals could generate data to predict or detect health risks or improve a person's treatment when necessary, it raises concern, as it permits near-constant surveillance and collection of excessive data that otherwise should remain unknown or uncollected. Such data collection also contributes to the ever-growing practice of "biosurveillance", a form of surveillance for health data and other biometrics, such as facial features, fingerprints, temperature and pulse (*32*). The growth of biosurveillance poses significant ethical and legal concerns, including the use of such data for medical and non-medical purposes for which explicit consent might not have been obtained or the repurposing of such data for non-health purposes by a government or company, such as within criminal justice or immigration systems. (See section 6.3.) Thus, such data should be liable to the same levels of data protection and security as for data collected on an individual in a formal clinical care setting.

Use of AI for resource allocation and prioritization

AI is being considered for use to assist in decision-making about prioritization or allocation of scarce resources. Prognostic scoring systems have long been available in critical care units. One of the best-known, Sequential Organ Failure Assessment (SOFA) (*33*), for analysis of the severity of illness and for predicting mortality, has been in use for decades, and SOFA scores have been widely used in some jurisdictions to guide allocation of resources for COVID-19 (*34*). It is not an AI system; however, an AI version, "DeepSOFA" (*35*), has been developed.

The growing attraction of this use of AI has been due partly to the COVID-19 pandemic, as many institutions lack bed capacity and others have inadequate ventilators. Thus, hospitals and clinics in the worst-affected countries have been overwhelmed.
It has been suggested that machine-learning algorithms could be trained and used to assist in decisions to ration supplies, identify which individuals should receive critical care or when to discontinue certain interventions, especially ventilator support (*36*).
AI tools could also be used to guide allocation of other scarce health resources during the COVID-19 pandemic, such as newly approved vaccines for which there is an insufficient initial supply (*37*).

Several ethical challenges associated with the use of AI for resource allocation and prioritization are described in section 6.5.

[1] Presentation by Christian Stammel. Wearable Technologies, Germany, to the WHO Meeting of the Expert Group on Ethics and Governance of AI for Health, 6 March 2020.

3.2 In health research and drug development

Application of AI for health research

An important area of health research with AI is based on use of data generated for electronic health records. Such data may be difficult to use if the underlying information technology system and database do not discourage the proliferation of heterogeneous or low-quality data. AI can nevertheless be applied to electronic health records for biomedical research, quality improvement and optimization of clinical care. From electronic health records, AI that is accurately designed and trained with appropriate data can help to identify clinical best practices before the customary pathway of scientific publication, guideline development and clinical support tools. AI can also assist in analysing clinical practice patterns derived from electronic health records to develop new clinical practice models.

A second (of many) application of AI for health research is in the field of genomics. Genomics is the study of the entire genetic material of an organism, which in humans consists of an estimated three billion DNA base pairs. Genomic medicine is an emerging discipline based on individuals' genomic information to guide clinical care and personalized approaches to diagnosis and treatment (*38*). As the analysis of such large datasets is complex, AI is expected to play an important role in genomics. In health research, for example, AI could improve human understanding of disease or identify new disease biomarkers (*38*), although the quality of the data and whether they are representative and unbiased (See section 6.6.) could undermine the results.

Uses of AI in drug development

AI is expected in time to be used to both simplify and accelerate drug development. AI could change drug discovery from a labour-intensive to a capital- and data-intensive process with the use of robotics and models of genetic targets, drugs, organs, diseases and their progression, pharmacokinetics, safety and efficacy. AI could be used in drug discovery and throughout drug development to shorten the process and make it less expensive and more effective (*39*). AI was used to identify potential treatments for Ebola virus disease, although, as in all drug development, identification of a lead compound may not result in a safe, effective therapy (*40*).

In December 2020, DeepMind announced that its AlphaFold system had solved what is known as the "protein folding problem", in that the system can reliably predict the three-dimensional shape of a protein (*41*). Although this achievement is only one part of a long process in understanding diseases and developing new medicines and vaccines, it should help to speed the development of new medicines and improve the repurposing of existing medicines for use against new viruses and new diseases (*41*). While this advance could significantly accelerate drug discovery, there is ethical concern about ownership and control of an AI technology that could be critical to drug development, as it might eventually be available to government, not-for-profit, academic and LMIC researchers only under commercial terms and conditions that limit its diffusion and use.

At present, drug development is led either by humans or by AI with human oversight. In the next two decades, as work with machines is optimized, the role of AI could evolve. Computing is starting to facilitate drug discovery and development by finding novel leads and evaluating whether they meet the criteria for new drugs, structuring unorganized data from medical imaging, searching large volumes of data, including health-care records, genetics data, laboratory tests, the Internet of Things, published literature and other types of health big data to identify structures and features, while recreating the body and its organs on chips (tissue chips) for AI analysis (*39*, *42*). By 2040, testing of medicines might be virtual – without animals or humans – based on computer models of the human body, tumours, safety, efficacy, epigenetics and other parameters. Prescription drugs could be designed for each person. Such efforts could contribute to precision medicine or health care that is individually tailored to a person's genes, lifestyle and environment.

3.3 In health systems management and planning

Health systems, even in a single-payer, government-run system, may be overly complex and involve numerous actors who contribute to, pay for or benefit from the provision of health-care services. The management and administration of care may be laborious. AI can be used to assist personnel in complex logistical tasks, such as optimization of the medical supply chain, to assume mundane, repetitive tasks or to support complex decision-making. Some possible functions of AI for health systems management include: identifying and eliminating fraud or waste, scheduling patients, predicting which patients are unlikely to attend a scheduled appointment and assisting in identification of staffing requirements (*43*).

AI could also be useful in complex decision-making and planning, including in LMIC. For example, researchers in South Africa applied machine-learning models to administrative data to predict the length of stay of health workers in underserved communities (*9*). In a study in Brazil, researchers used several government data sets and AI to optimize the allocation of health-system resources by geographical location according to current health challenges (*9*). Allocation of scarce health resources through use of AI has raised concern, however, that resources may not be fairly allocated due, for example, to bias in the data. (See section 6.5.)

3.4 In public health and public health surveillance

Several AI tools for population and public health can be used in public health programmes. For example, new developments in AI could, after rigorous evaluation, improve identification of disease outbreaks and support surveillance. Several concerns about the use of technology for public health surveillance, promotion and outbreak response must, however, be considered before use of AI for such purposes, including the tension between the public health benefits of surveillance and ethical and legal concern about individual (or community) privacy and autonomy (*44*).

Health promotion

AI can be used for health promotion or to identify target populations or locations with "high-risk" behaviour and populations that would benefit from health communication and messaging (micro-targeting). AI programmes can use different forms of data to identify such populations, with varying accuracy, to improve message targeting. Micro-targeting can also, however, raise concern, such as that with respect to commercial and political advertising, including the opaqueness of processes that facilitate micro-targeting. Furthermore, users who receive such messages may have no explanation or indication of why they have been targeted (*45*). Micro-targeting also undermines a population's equal access to information, can affect public debate and can facilitate exclusion or discrimination if it is used improperly by the public or private sector.

Disease prevention

AI has also been used to address the underlying causes of poor health outcomes, such as risks related to environmental or occupational health. AI tools can be used to identify bacterial contamination in water treatment plants, simplify detection and lower the costs. Sensors can also be used to improve environmental health, such as by analysing air pollution patterns or using machine learning to make inferences between the physical environment and healthy behaviour (*29*). One concern with such use of AI is whether it is provided equitably or if such technologies are used only on behalf of wealthier populations and regions that have the relevant infrastructure for its use (*46*).

Surveillance (including prediction-based surveillance) and emergency preparedness

AI has been used in public health surveillance for collecting evidence and using it to create mathematical models to make decisions. Technology is changing the types of data collected for public health surveillance by the addition of digital "traces", which are data that are not generated specifically for public health purposes (such as from blogs, videos, official reports and Internet searches). Videos (e.g. YouTube) are another "rich" source of information for health insights (*47*).

Characterization of digital traces as "health data" raises questions about the types of privacy protection or other safeguards that should be attached to such datasets if they are not publicly available. For example, the use of digital traces as health data could violate the data protection principle of "purpose limitation", that individuals who generate such data should know what their data will be used for at the point of collection (*48*).

Such use also raises questions of accuracy. Models are useful only when appropriate data are used. Machine-learning algorithms could be more valuable when augmented by digital traces of human activity, yet such digital traces could also negatively impact an algorithm's performance. Google Flu Trends, for example, was based on search engine queries about complications, remedies, symptoms and antiviral medications for

influenza, which are used to estimate and predict influenza activity. While Google Flu Trends first provided relatively accurate predictions before those of the US Centers for Disease Control and Prevention, it overestimated the prevalence of flu between 2011 and 2013 because the system was not re-trained as human search behaviour evolved (*49*).

Although many public health institutions are not yet making full use of these sources of data, surveillance itself is changing, especially real-time surveillance. For example, researchers could detect a surge in cases of severe pulmonary disease associated with the use of electronic cigarettes by mining disparate online sources of information and using Health Map, an online data-mining tool (*50*). Similarly, Microsoft researchers have found early evidence of adverse drug reactions from web logs with an AI system. In 2013, the company's researchers detected side-effects of several prescription drugs before they were found by the US Food and Drug Administration's warning system (*51*). In 2020, the US Food and Drug Administration sponsored a "challenge", soliciting public submissions to develop computation algorithms for automatic detection of adverse events from publicly available data *(52*). Despite its potential benefits, real-time data collection, like the collection and use of digital traces, could violate data protection rules if surveillance was not the purpose of its initial collection, which is especially likely when data collection is automated.

Before the COVID-19 pandemic, WHO had started to develop EPI-BRAIN, a global platform that will allow experts in data and public health to analyse large datasets for emergency preparedness and response. (See also section 7.1.) AI has been used to assist in both detection and prediction during the COVID-19 pandemic, although some consider that the techniques and programming developed will "pay dividends" only during a subsequent pandemic (*49*). HealthMap first issued a short bulletin about a new type of pneumonia in Wuhan, China, at the end of December 2019 (*49*). Since then, AI has been used to "now-cast" (assess the current state of) the COVID-19 pandemic (*49*), while, in some countries, real-time data on the movement and location of people has been used to build AI models to forecast regional transmission dynamics and guide border checks and surveillance (*53*). In order to determine how such applications should be used, an assessment should be conducted of whether they are accurate, effective and useful.

Outbreak response

The possible uses of AI for different aspects of outbreak response have also expanded during the COVID-19 pandemic. They include studying SARS-CoV2 transmission, facilitating detection, developing possible vaccines and treatments and understanding the socio-economic impacts of the pandemic (*54*). Such use of AI was already tested during the pandemic of Ebola virus disease in West Africa in 2014, although the assumptions underlying use of AI technologies to predict the spread of the Ebola virus were based on erroneous views of how the virus was spreading (*55*, *56*). While many

possible uses of AI have been identified and used during the COVID-19 pandemic, their actual impact is likely to have been modest; in some cases, early AI screening tools for SARS-CoV2 "were utter junk" with which companies "were trying to capitalise on the panic and anxiety" (*57*).

New applications (*58*) are intended to support the off-line response, although not all may involve use of AI. These have included proximity tracking applications intended to notify users (and possibly health authorities) that they have been in the proximity (for some duration) of an individual who subsequently tested positive for SARS-CoV2. Concern has been raised about privacy and the utility and accuracy of proximity-tracking applications, and WHO issued interim guidance on the ethical use of proximity-tracking applications in 2020 (*59*).

WHO and many ministries of health have also deployed symptom checkers, which are intended to guide users through a series of questions to assist in determining whether they should seek additional medical advice or testing for SARS-CoV2. The first symptom checkers were "hard coded", based on accumulated clinical judgement, as there were no previous data, and on a simple decision tree from older AI techniques, which involved direct encoding of expert knowledge. AI systems based on machine learning require accurate training, while data are initially scarce for a new disease such as COVID-19 (*60*). New symptom checkers are based on machine learning to provide advice to patients (*61*), although their effectiveness is not yet known; all symptom checkers require that users provide accurate information.

AI has also been introduced to map the movements of individuals in order to approximate the effectiveness of government-mandated orders to remain in confinement, and, in some countries, AI technology has been used to identify individuals who should self-quarantine and be tested. These technologies raise legal and ethical concerns about privacy and risk of discrimination and also about possibly unnecessary restriction of movement or access to services, which heavily impact the exercise of a range of human rights (*53*). As for all AI technologies, their actual effectiveness depends on whether the datasets are representative of the populations in which the technologies are used, and they remain questionable without systematic testing and evaluation. The uses described above are therefore not yet established.

3.5 The future of artificial intelligence for health

While AI may not replace clinical decision-making, it could improve decisions made by clinicians. In settings with limited resources, AI could be used to conduct screening and evaluation if insufficient medical expertise is available, a common challenge in many resource-poor settings. Yet, whether AI can advance beyond narrow tasks depends on numerous factors beyond the state of AI science and on the trust of providers, patients and health-care professionals in AI-based technologies. In the

following sections of this report, ethical concerns and risks associated with the expanding use of AI for health are discussed, including by whom and how such technologies are deployed and developed. Technological, legal, security and ethical challenges and concerns are discussed not to dissuade potential use of AI for health but to ensure that AI fulfils its great potential and promise.

4. LAWS, POLICIES AND PRINCIPLES THAT APPLY TO ARTIFICIAL INTELLIGENCE FOR HEALTH

Laws, policies and principles for regulating and managing the use of AI and specifically use of AI for health are fragmented and limited. Numerous principles and guidelines have been developed for application of "ethical" AI in the private and public sectors and in research institutions (*62*); however, there is no consensus on its definition, best practices or ethical requirements, and different legal regimes and governance models are associated with each set of principles. Other norms, rules and frameworks also apply to use of AI, including human rights obligations, bioethics laws and policies, data protection laws and regulatory standards. These are summarized below and discussed elsewhere in the report. Section 5 provides a set of guiding principles agreed by the WHO Expert Group by consensus, on which this analysis and these findings are based.

4.1 Artificial intelligence and human rights

Efforts to enumerate human rights and to fortify their observance through explicit legal mechanisms are reflected in international and regional human rights conventions, including the Universal Declaration on Human Rights, the International Covenant on Economic, Social and Cultural Rights (including General Comment No. 14, which defines the right to health), the International Covenant on Civil and Political Rights and regional human rights conventions, such as the African Charter on Human and People's Rights, the American Convention on Human Rights and the European Convention on Human Rights. Not all governments have acceded to key human rights instruments; some have signed but not ratified such charters or have expressed reservations to certain provisions. In general, however, human rights listed in international instruments establish a baseline for the protection and promotion of human dignity worldwide and are enforced through national legislation such as constitutions or human rights legislation.

Machine-learning systems could advance human rights but could also undermine core human rights standards. The Office of the High Commissioner for Human Rights has issued several opinions on the relation of AI to the realization of human rights. In guidance issued in March 2020, the Office noted that AI and big data can improve the human right to health when "new technologies are designed in an accountable manner" and could ensure that certain vulnerable populations have efficient, individualized care, such as assistive devices, built-in environmental applications and robotics (*63*). The Office also noted, however, that such technologies could dehumanize care, undermine the autonomy and independence of older persons and pose significant risks to patient privacy – all of which are contrary to the right to health (*63*). In February 2021, in a speech to the Human Rights Council, the United Nations

Secretary-General noted a number of concerns for human rights associated with the growing collection and use of data on the COVID-19 pandemic and called on governments to "place human rights at the centre of regulatory frameworks and legislation on the development and use of digital technologies" (*64*). Human rights organizations have interpreted and, when necessary, adapted existing human rights laws and standards to AI assessment and are reviewing them in the face of the challenges and opportunities associated with AI. The Toronto Declaration (*65*) addresses the impact of AI on human rights and situates AI within the universally binding, actionable framework of human rights laws and standards; it provides mechanisms for public and private sector accountability and the protection of people from discrimination and promotes equity, diversity and inclusion, while safeguarding equality and effective redress and remedy.

In 2018, the Council of Europe's Committee of Ministers issued draft recommendations to Member States on the impact of algorithmic systems on human rights (*66*). The Council of Europe is further examining the feasibility and potential elements of a legal framework for the development, design and application of digital technologies according to its standards on human rights, democracy and the rule of law.
Legal frameworks for human rights, bioethics and privacy adopted by countries are applicable to several aspects of AI for health. They include Article 8 of the European Convention on Human Rights: the right to respect for private and family life, home and correspondence (*67*); the Oviedo Convention on Human Rights and Biomedicine, which covers ethical principles of individual human rights and responsibilities (68); the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (*69*) and guidelines on the protection of individuals with regard to the processing of personal data in a world of big data, prepared by the Consultative Committee of Convention 108+ (*69*).

Yet, even with robust human rights standards, organizations and institutions recognize that better definition is required of how human rights standards and safeguards relate and apply to AI and that new laws and jurisprudence are required to address the interaction of AI and human rights. New legal guidance has been prepared by the Council of Europe. In 2019–2020, the Council established the Ad-hoc Committee on Artificial Intelligence to conduct broad multi-stakeholder consultations in order to determine the feasibility and potential elements of a legal framework for the design and application of AI according to the Council of Europe's standards on human rights, democracy and the rule of law. Further, in 2019, the Council of Europe released Guidelines on artificial intelligence and data protection (*70*), also based on the protection of human dignity and safeguarding human rights and fundamental freedom. In addition, the ethical charter of the European Commission for Efficiency of Justice includes five principles relevant to use of AI for health (*71*).

4.2 Data protection laws and policies

Data protection laws are "rights-based approaches" that provide standards for regulating data processing that both protect the rights of individuals and establish obligations for data controllers and processors. Data protection laws also increasingly recognize that people have the right not to be subject to decisions guided solely by automated processes. Over 100 countries have enacted data protection laws. One well-known set of data protection laws is the General Data Protection Regulation (GDPR) of the European Union (EU); in the USA, the Health Insurance Portability and Accountability Act, enacted in 1996, applies to privacy and to the security of health data.

Some standards and guidelines are designed specifically to manage the use of personal data for AI. For example, the Ibero-American Data Protection Network, which consists of 22 data protection authorities in Portugal and Spain and in Mexico and other countries in Central and South America and the Caribbean, has issued General Recommendations for the Processing of Personal Data in Artificial Intelligence (72) and specific guidelines for compliance with the principles and rights that govern the protection of personal data in AI projects (*73*).

4.3 Existing laws and policies related to health data

Several types of laws and policies govern the collection, processing, analysis, transfer and use of health data. The Council of Europe's Committee of Ministers issued a recommendation to Member States on the protection of health-related data in 2019 (*74*), and the African Union's convention on cybersecurity and personal data protection (2014) (*75*) requires that personal data involving genetic information and health research be processed only with the authorization of the national data protection authority through the Personal Data Protection Guidelines for Africa (*76*). Generally, the African continent's digital transformation strategy (*77*) encourages African Union Member States to "have adequate regulation; particularly around data governance and digital platforms, to ensure that trust is preserved in the digitalization". In February 2021, the African Academy of Sciences and the African Union Development Agency released recommendations for data and biospecimen governance in Africa to promote a participant-centred approach to research involving human participants, while enabling ethical research practices on the continent and providing guidelines for governance (*78*).

Laws that govern the transfer of data among countries include those defined in trade agreements, intellectual property (IP) rules for the ownership of data and the role of competition law and policy related to the accumulation and control of data (including health data). These are discussed in detail later in this report.

4.4 General principles for the development and use of artificial intelligence

An estimated 100 proposals for AI principles have been published in the past decade, and studies have been conducted to identify which principles are most cited (*79*). In one study of mapping and analysis of current principles and guidelines for ethical use of AI, convergence was found on transparency, justice, fairness, non-maleficence and responsibility, while other principles such as privacy, solidarity, human dignity and sustainability were under-represented (*62*).

Several intergovernmental organizations and countries have proposed such principles (Box 1).

Box 1. Examples of AI ethics principles proposed by intergovernmental organizations and countries

- The Recommendations of the OECD Council on Artificial Intelligence (*80*), the first intergovernmental standard on AI, were adopted in May 2019 by OECD's 36 member countries and have since been applied by a number of partner economies. The OECD AI principles (*81*) provided the basis for the AI principles endorsed by G20 governments in June 2019 (*82*). While OECD recommendations are not legally binding, they carry a political commitment and have proved highly influential in setting international standards in other policy areas (e.g. privacy and data protection) and helping governments to design national legislation. The OECD launched an online platform for public policy on AI, the AI Policy Observatory (*83*) (See section 9.6.) and is cooperating on this and other initiatives on the ethical implications of AI with the Council of Europe, the United Nations Economic, Scientific and Cultural Organization (UNESCO) and WHO.
- In 2019, the Council of Europe Commissioner for Human Rights issued recommendations to ensure that human rights are strengthened rather than undermined by AI: Unboxing artificial intelligence: 10 steps to protect human rights recommendations (*84*).
- The European Commission appointed 52 representatives from academia, civil society and industry to its High-level Expert Group on Artificial Intelligence and issued Ethics Guidelines for Trustworthy AI (*85*).
- Japan has issued several guidelines on the use of AI, including on research and development and utilization (*86*).
- China has issued National Governance Principles for the New Generation Artificial Intelligence, which serves as the national principles for AI governance in China (*87*). Academia and industry have jointly issued the Beijing Artificial Intelligence Principles (*88*).[2]
- In Singapore, a series of initiatives on AI governance and ethics was designed to build an ecosystem of trust to support adoption of AI. They include Asia's first Model AI governance framework, released in January 2019; an international industry-led Advisory Council on the Ethical Use of AI and Data formed in June 2018; a research programme on the governance of AI and data use established in partnership with the Singapore Management University in September 2018 (*89*); and a certification programme for ethics and governance of AI for companies and developers (*90*).
- The African Union's High-level Panel on Emerging Technologies is preparing broad guidance on the use of AI to promote economic development and its use in various sectors, including health care (*91*).

[2] Presentation by Professor Yi Zeng, Chinese Academy of Sciences, 4 October 2019, to the WHO working group on ethics and governance of AI for health.

4.5 Principles for use of artificial intelligence for health

No specific ethical principles for use of AI for health have yet been proposed for adoption worldwide. Before WHO's work on guidance on the ethics and governance of AI for health, the WHO Global Conference on Primary Health Care issued the Astana Declaration (*92*), which includes principles for the use of digital technology. The Declaration calls for promotion of rational, safe use and protection of personal data and use of technology to improve access to health care, enrich health service delivery, improve the quality of service and patient safety and increase the efficiency and coordination of care.

UNESCO has guidance and principles for the use of AI in general and for the use of big data in health. UNESCO's work on the ethical implications of AI is supported by two standing expert committees, the World Commission on the Ethics of Scientific Knowledge and Technology and the International Bioethics Committee. Other work includes the report of the International Bioethics Committee on big data and health in 2017, which identified important elements of a governance framework (*93*); the World Commission on the Ethics of Scientific Knowledge and Technology report on robotics ethics in 2017 (*94*); a preliminary study on the ethics of AI by UNESCO in 2019, which raised ethical concern about education, science and gender (*95*); a recommendation on the ethics of AI to be considered by UNESCO's General Conference in 2021; and a report by the World Commission on the Ethics of Scientific Knowledge and Technology on the Internet of Things.

In 2019, the United Kingdom's National Health Service (NHS) released a code of conduct, with 10 principles for the development and use of safe, ethical, effective, data-based health and care technologies (*96*). In October 2019, The Lancet and The Financial Times launched a joint commission, The Governing Health Futures 2030: Growing up in a Digital World Commission, on the convergence of digital health, AI and universal health coverage, which will consult between October 2019 and December 2021 (*97*).

4.6 Bioethics laws and policies

Bioethics laws and policies play a role in regulating the use of AI, and several bioethics laws have been revised in recent years to include recognition of the growing use of AI in science, health care and medicine. The French Government's most recent revision of its national bioethics law (98), which was endorsed in 2019, establishes standards to address the rapid growth of digital technologies in the health-care system. It includes standards for human supervision, or human warranty, that require evaluation by patients and clinicians at critical points in the development and deployment of AI. It also supports free, informed consent for the use of data and the creation of a secure national platform for the collection and processing of health data.

4.7 Regulatory considerations

Regulation of AI technologies is likely to be developed and implemented by health regulatory authorities responsible for ensuring the safety, efficacy and appropriate use of technologies for health care and therapeutic development. A WHO expert group that is preparing considerations for the regulation of AI for health has discussed areas that should be considered by stakeholders, including developers and regulators, in examining new AI technologies. They include documentation and transparency, risk management and the life-cycle approach, data quality, analytical and clinical validation, engagement and collaboration, and privacy and data protection. Many regulatory authorities are preparing considerations and frameworks for the use of AI, and they should be examined, potentially with the relevant regulatory agency. Governance of AI through regulatory frameworks and the ethical principles that should be considered are discussed in section 9.5.

5. KEY ETHICAL PRINCIPLES FOR USE OF ARTIFICIAL INTELLIGENCE FOR HEALTH

Ethical principles for the application of AI for health and other domains are intended to guide developers, users and regulators in improving and overseeing the design and use of such technologies. Human dignity and the inherent worth of humans are the central values upon which all other ethical principles rest.

An ethical principle is a statement of a duty or a responsibility in the context of the development, deployment and continuing assessment of AI technologies for health. The ethical principles described below are grounded in basic ethical requirements that apply to all persons and that are considered noncontroversial. The requirements are as follows.

- Avoid harming others (sometimes called "Do no harm" or nonmaleficence).
- Promote the well-being of others when possible (sometimes called "beneficence"). Risks of harm should be minimized, while maximizing benefits. Expected risks should be balanced against expected benefits.
- Ensure that all persons are treated fairly, which includes the requirement to ensure that no person or group is subject to discrimination, neglect, manipulation, domination or abuse (sometimes called "justice" or "fairness").
- Deal with persons in ways that respect their interests in making decisions about their lives and their person, including health-care decisions, according to informed understanding of the nature of the choice to be made, its significance, the person's interests and the likely consequences of the alternatives (sometimes called "respect for persons" or "autonomy").

Additional moral requirements can be derived from this list of fundamental moral requirements. For example, safeguarding and protecting individual privacy is not only recognized as a legal requirement in many countries but is also important to enable people to control sensitive information about themselves and self-determination (respect for their autonomy) and to avoid harm.

These ethical principles are intended to provide guidance to stakeholders about how basic moral requirements should direct or constrain their decisions and actions in the specific context of developing, deploying and assessing the performance of AI technologies for health. These principles are also intended to emphasize issues that arise from the use of a technology that could alter relations of moral significance. For example, it has long been recognized that health-care providers have a special duty to advance these values with respect to patients because of the centrality of health to

individual well-being, because of the dependence of patients on health professionals for information about their diagnosis, prognosis and the relative merits of the available treatment or prevention options, and the importance of free and open exchange of information to the provider–patient relationship. If AI systems are used by health-care workers to conduct clinical tasks or to delegate clinical tasks that were once reserved for humans, programmers who design and program such AI technologies should also adhere to these ethical obligations.

Thus, the ethical principles are important for all stakeholders who seek guidance in the responsible development, deployment and evaluation of AI technologies for health, including clinicians, systems developers, health system administrators, policy-makers in health authorities, and local and national governments. The ethical principles listed here should encourage and assist governments and public sector agencies to keep pace with the rapid evolution of AI technologies through legislation and regulation and should empower medical professionals to use AI technologies appropriately.
Ethical principles should also be embedded within professional and technological standards for AI. Software engineers already are guided by standards such as for fitness for purpose, documentation and provenance, and version control. Standards are required to guide the interoperability and design of a program, for continuing education of those who develop and use such technologies and for governance. Moreover, the standards for the evaluation and external audit of systems are evolving in the context of their use. In health computing, there are standards for system integration, electronic health records, system interoperability, implementation and programming structures.

Although ethical principles do not always clearly address limitations in the uses of such technologies, governments should ban or restrict the use of AI or other technologies if they violate or imperil the exercise of human rights, do not conform to other principles or regulations or would be introduced in unprepared or other inappropriate contexts. For example, many countries lack data protection laws or have inadequate regulatory frameworks to guide the introduction of AI technologies.

The claim that certain basic moral requirements must constrain and guide the conduct of persons can also be expressed in the language of human rights. Human rights are intended to capture a basic set of moral and legal requirements for conduct to which every person is entitled regardless of race, sex, nationality, ethnicity, language, religion or any other feature. These rights include human dignity, equality, non-discrimination, privacy, freedom, participation, solidarity and accountability.

Machine-learning systems could advance the protection and enforcement of human rights (including the human right to health) but could undermine core human rights such as non-discrimination and privacy. Human rights and ethical principles are intimately interlinked; because human rights are legally binding, they provide a

powerful framework by which governments, international organizations and private actors are obligated to abide. Private sector actors have the responsibility to respect human rights, independently of state obligations. In fulfilling this responsibility, private sector actors must take continuous proactive and reactive steps to ensure that they do not abuse or contribute to the abuse of human rights.

The existence of a human rights framework does not, however, obviate the need for continuing ethical deliberation. Indeed, much of ethics is intended to expand upon and complement the norms and obligations established in human rights agreements. In many situations, multiple ethical considerations are relevant and require weighing up and balancing to accommodate the multiple principles at stake. An ethically acceptable decision depends on consideration of the full range of appropriate ethical considerations, ensuring that multiple perspectives are factored into the analysis and creating a decision-making process that stakeholders will consider fair and legitimate.

This guidance identifies six ethical principles to guide the development and use of AI technology for health. While ethical principles are universal, their implementation may differ according to the cultural, religious and other social context. Many of the ethical issues arising in the use of AI and machine learning are not completely new but have arisen for other applications of information and communication technologies for health, such as use of any computer to track a disease or make a diagnosis or prognosis. Computers were performing these tasks with various programs long before AI became noteworthy. Ethical guidance and related principles have been articulated for fields such as telemedicine and data-sharing. Likewise, several ethical frameworks have been developed for AI in general, outside the health sector. (See section 4.) The ethical principles listed here are those identified by the WHO Expert Group as the most appropriate for the use of AI for health.

5.1 Protect autonomy

Adoption of AI can lead to situations in which decision-making could be or is in fact transferred to machines. The principle of autonomy requires that any extension of machine autonomy not undermine human autonomy. In the context of health care, this means that humans should remain in full control of health-care systems and medical decisions. AI systems should be designed demonstrably and systematically to conform to the principles and human rights with which they cohere; more specifically, they should be designed to assist humans, whether they be medical providers or patients, in making informed decisions. Human oversight may depend on the risks associated with an AI system but should always be meaningful and should thus include effective, transparent monitoring of human values and moral considerations. In practice, this could include deciding whether to use an AI system for a particular health-care decision, to vary the level of human discretion and decision-making and to develop AI technologies that can rank decisions when appropriate (as opposed to a single decision). These practices

can ensure a clinician can override decisions made by AI systems and that machine autonomy can be restricted and made "intrinsically reversible".

Respect for autonomy also entails the related duties to protect privacy and confidentiality and to ensure informed, valid consent by adopting appropriate legal frameworks for data protection. These should be fully supported and enforced by governments and respected by companies and their system designers, programmers, database creators and others. AI technologies should not be used for experimentation or manipulation of humans in a health-care system without valid informed consent. The use of machine-learning algorithms in diagnosis, prognosis and treatment plans should be incorporated into the process for informed and valid consent. Essential services should not be circumscribed or denied if an individual withholds consent and that additional incentives or inducements should not be offered by either a government or private parties to individuals who do provide consent.

Data protection laws are one means of safeguarding individual rights and place obligations on data controllers and data processors. Such laws are necessary to protect privacy and the confidentiality of patient data and to establish patients' control over their data. Construed broadly, data protection laws should also make it easy for people to access their own health data and to move or share those data as they like. Because machine learning requires large amounts of data – big data – these laws are increasingly important.

5.2 Promote human well-being, human safety and the public interest

AI technologies should not harm people. They should satisfy regulatory requirements for safety, accuracy and efficacy before deployment, and measures should be in place to ensure quality control and quality improvement. Thus, funders, developers and users have a continuous duty to measure and monitor the performance of AI algorithms to ensure that AI technologies work as designed and to assess whether they have any detrimental impact on individual patients or groups.

Preventing harm requires that use of AI technologies does not result in any mental or physical harm. AI technologies that provide a diagnosis or warning that an individual cannot address because of lack of appropriate, accessible or affordable health care should be carefully managed and balanced against any "duty to warn" that might arise from incidental and other findings, and appropriate safeguards should be in place to protect individuals from stigmatization or discrimination due to their health status.

5.3 Ensure transparency, explainability and intelligibility

AI should be intelligible or understandable to developers, users and regulators. Two broad approaches to ensuring intelligibility are improving the transparency and explainability of AI technology.

Transparency requires that sufficient information (described below) be published or documented before the design and deployment of an AI technology. Such information should facilitate meaningful public consultation and debate on how the AI technology is designed and how it should be used. Such information should continue to be published and documented regularly and in a timely manner after an AI technology is approved for use.

Transparency will improve system quality and protect patient and public health safety. For instance, system evaluators require transparency in order to identify errors, and government regulators rely on transparency to conduct proper, effective oversight. It must be possible to audit an AI technology, including if something goes wrong. Transparency should include accurate information about the assumptions and limitations of the technology, operating protocols, the properties of the data (including methods of data collection, processing and labelling) and development of the algorithmic model.

AI technologies should be explainable to the extent possible and according to the capacity of those to whom the explanation is directed. Data protection laws already create specific obligations of explainability for automated decision-making. Those who might request or require an explanation should be well informed, and the educational information must be tailored to each population, including, for example, marginalized populations. Many AI technologies are complex, and the complexity might frustrate both the explainer and the person receiving the explanation. There is a possible trade-off between full explainability of an algorithm (at the cost of accuracy) and improved accuracy (at the cost of explainability).

All algorithms should be tested rigorously in the settings in which the technology will be used in order to ensure that it meets standards of safety and efficacy. The examination and validation should include the assumptions, operational protocols, data properties and output decisions of the AI technology. Tests and evaluations should be regular, transparent and of sufficient breadth to cover differences in the performance of the algorithm according to race, ethnicity, gender, age and other relevant human characteristics. There should be robust, independent oversight of such tests and evaluation to ensure that they are conducted safely and effectively.

Health-care institutions, health systems and public health agencies should regularly publish information about how decisions have been made for adoption of an AI technology and how the technology will be evaluated periodically, its uses, its known limitations and the role of decision-making, which can facilitate external auditing and oversight.

5.4 Foster responsibility and accountability

Humans require clear, transparent specification of the tasks that systems can perform and the conditions under which they can achieve the desired level of performance; this helps to ensure that health-care providers can use an AI technology responsibly. Although AI technologies perform specific tasks, it is the responsibility of human stakeholders to ensure that they can perform those tasks and that they are used under appropriate conditions.

Responsibility can be assured by application of "human warranty", which implies evaluation by patients and clinicians in the development and deployment of AI technologies. In human warranty, regulatory principles are applied upstream and downstream of the algorithm by establishing points of human supervision. The critical points of supervision are identified by discussions among professionals, patients and designers. The goal is to ensure that the algorithm remains on a machine-learning development path that is medically effective, can be interrogated and is ethically responsible; it involves active partnership with patients and the public, such as meaningful public consultation and debate (*101*). Ultimately, such work should be validated by regulatory agencies or other supervisory authorities.

When something does go wrong in application of an AI technology, there should be accountability. Appropriate mechanisms should be adopted to ensure questioning by and redress for individuals and groups adversely affected by algorithmically informed decisions. This should include access to prompt, effective remedies and redress from governments and companies that deploy AI technologies for health care. Redress should include compensation, rehabilitation, restitution, sanctions where necessary and a guarantee of non-repetition.

The use of AI technologies in medicine requires attribution of responsibility within complex systems in which responsibility is distributed among numerous agents. When medical decisions by AI technologies harm individuals, responsibility and accountability processes should clearly identify the relative roles of manufacturers and clinical users in the harm. This is an evolving challenge and remains unsettled in the laws of most countries. Institutions have not only legal liability but also a duty to assume responsibility for decisions made by the algorithms they use, even if it is not feasible to explain in detail how the algorithms produce their results.

To avoid diffusion of responsibility, in which "everybody's problem becomes nobody's responsibility", a faultless responsibility model ("collective responsibility"), in which all the agents involved in the development and deployment of an AI technology are held responsible, can encourage all actors to act with integrity and minimize harm. In such a model, the actual intentions of each agent (or actor) or their ability to control an outcome are not considered.

5.5 Ensure inclusiveness and equity

Inclusiveness requires that AI used in health care is designed to encourage the widest possible appropriate, equitable use and access, irrespective of age, gender, income, ability or other characteristics. Institutions (e.g. companies, regulatory agencies, health systems) should hire employees from diverse backgrounds, cultures and disciplines to develop, monitor and deploy AI. AI technologies should be designed by and evaluated with the active participation of those who are required to use the system or will be affected by it, including providers and patients, and such participants should be sufficiently diverse. Participation can also be improved by adopting open-source software or making source codes publicly available.

AI technology – like any other technology – should be shared as widely as possible. AI technologies should be available not only in HIC and for use in contexts and for needs that apply to high-income settings but they should also be adaptable to the types of devices, telecommunications infrastructure and data transfer capacity in LMIC. AI developers and vendors should also consider the diversity of languages, ability and forms of communication around the world to avoid barriers to use. Industry and governments should strive to ensure that the "digital divide" within and between countries is not widened and ensure equitable access to novel AI technologies.
AI technologies should not be biased. Bias is a threat to inclusiveness and equity because it represents a departure, often arbitrary, from equal treatment. For example, a system designed to diagnose cancerous skin lesions that is trained with data on one skin colour may not generate accurate results for patients with a different skin colour, increasing the risk to their health.

Unintended biases that may emerge with AI should be avoided or identified and mitigated. AI developers should be aware of the possible biases in their design, implementation and use and the potential harm that biases can cause to individuals and society. These parties also have a duty to address potential bias and avoid introducing or exacerbating health-care disparities, including when testing or deploying new AI technologies in vulnerable populations.

AI developers should ensure that AI data, and especially training data, do not include sampling bias and are therefore accurate, complete and diverse. If a particular racial or ethnic minority (or other group) is underrepresented in a dataset, oversampling of that group relative to its population size may be necessary to ensure that an AI technology achieves the same quality of results in that population as in better-represented groups.

AI technologies should minimize inevitable power disparities between providers and patients or between companies that create and deploy AI technologies and those that use or rely on them. Public sector agencies should have control over the data collected

by private health-care providers, and their shared responsibilities should be defined and respected. Everyone – patients, health-care providers and health-care systems – should be able to benefit from an AI technology and not just the technology providers. AI technologies should be accompanied by means to provide patients with knowledge and skills to better understand their health status and to communicate effectively with health-care providers. Future health literacy should include an element of information technology literacy.

The effects of use of AI technologies must be monitored and evaluated, including disproportionate effects on specific groups of people when they mirror or exacerbate existing forms of bias and discrimination. Special provision should be made to protect the rights and welfare of vulnerable persons, with mechanisms for redress if such bias and discrimination emerges or is alleged.

5.6 Promote artificial intelligence that is responsive and sustainable

Responsiveness requires that designers, developers and users continuously, systematically and transparently examine an AI technology to determine whether it is responding adequately, appropriately and according to communicated expectations and requirements in the context in which it is used. Thus, identification of a health need requires that institutions and governments respond to that need and its context with appropriate technologies with the aim of achieving the public interest in health protection and promotion. When an AI technology is ineffective or engenders dissatisfaction, the duty to be responsive requires an institutional process to resolve the problem, which may include terminating use of the technology.

Responsiveness also requires that AI technologies be consistent with wider efforts to promote health systems and environmental and workplace sustainability. AI technologies should be introduced only if they can be fully integrated and sustained in the health-care system. Too often, especially in under-resourced health systems, new technologies are not used or are not repaired or updated, thereby wasting scare resources that could have been invested in proven interventions. Furthermore, AI systems should be designed to minimize their ecological footprints and increase energy efficiency, so that use of AI is consistent with society's efforts to reduce the impact of human beings on the earth's environment, ecosystems and climate. Sustainability also requires governments and companies to address anticipated disruptions to the workplace, including training of health-care workers to adapt to use of AI and potential job losses due to the use of automated systems for routine health-care functions and administrative tasks.