TRUSTWORTHY AI IN HEALTH

Background paper for the G20 AI Dialogue, Digital Economy Task Force

SAUDI ARABIA,
1-2 APRIL 2020

OECD
BETTER POLICIES FOR BETTER LIVES

This document was prepared by the Organisation for Economic Co-operation and Development's Directorate for Employment, Labour and Social Affairs, and Directorate for Science, Technology and Innovation, as an input for the discussions in the G20 Digital Economy Task Force in 2020, under the auspices of the Saudi Arabia Presidency in 2020. The opinions expressed and arguments employed herein do not necessarily represent the official views of the member countries of the OECD or the G20.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Cover image: Jason Leung on Unsplash.

In Brief

Key messages

Artificial intelligence (AI) is talked about as potentially delivering a **genuine leap in global productivity** – a pillar of the 4th industrial revolution – with an impact on humanity (and the planet) as profound as those of steam power and electricity. There is great potential to generate benefits to human welfare but the risks of harm are equally high. Like any powerful technology, AI can serve to entrench and amplify existing socio-economic problems, rather than ameliorating them. At worst, it can be deployed towards nefarious and destructive purposes. Which path will be taken will rely on political and policy choices.

The **potential of AI in health is profound**, given the growing volume of electronic data as well as the inherent complexity of the sector, its reliance on information to solve problems, and the variability and complexity of how disease interacts with individuals and populations. AI is a 'general purpose' technology that can be deployed in just about any facet or activity of the health industry, from clinical decision-making and public health, to biomedical research and drug development, to health system administration and service redesign. As the COVID-19 crisis is showing, there are genuine opportunities for AI to deliver benefits for health systems, professionals and the public, making existing clinical and administrative processes more effective, efficient and equitable. In a notoriously wasteful and inefficient industry, this is a major opportunity to improve health outcomes and value for money.

The risks of **unintended and negative consequences** associated with AI are commensurately high, especially at scale. Most AI in health is actually artificial narrow intelligence, designed to accomplish a very specific task on previously curated data from one setting. In the real world, health data are unstandardised, patient populations are diverse, and biased decision-makers make mistakes that are then reflected in data. Because most AI models build on correlations, predictions could fail to generalise to different populations or settings, and might exacerbate existing inequalities and biases. As the AI industry is extremely gender and race imbalanced, and health professionals are already overwhelmed by other digital tools, there could be little capacity to catch errors and resistance from clinicians.

Policy makers should **beware the hype and focus on real problems** and opportunities. One key priority is to improve data quality, interoperability and access in a secure way through better data governance. Overarching this is a need to operationalise the G20 AI Principles – both the values-based principles for responsible stewardship of trustworthy AI and the recommendations for national policies and international co-operation – in health through industry-specific policy and regulatory reform. This may have a focus on, among other things, ensuring transparency and explainability of AI, accountability for the outputs of AI models, regulatory oversight, building capacity among health workers (and the public), and long-term investment. Strong policy frameworks based on inclusive and extensive dialogue among all stakeholders are needed to **ensure AI adds value to patients and to society**. This includes balancing the prospective benefits and risks of private sector AI initiatives in health, where profit motives may clash with the public interest. Regulatory sandboxes are a useful way to test the utility and scalability of AI while ring-fencing wider health systems from risks.

There is no indication that machines will replace humans who work in health, but AI will almost definitely result in human tasks changing greatly. AI that influences clinical and public health decisions should be introduced with care. **High expectations must be managed, but real opportunities should be pursued**. Most AI in health is far from the versatile general learning humans are capable of, focussing only on a specific, predefined task by using data curated by humans. Nonetheless, there is little doubt

change is coming, and that some processes, workflows and institutions will need to adapt to make the most of the opportunities offered by AI, while managing the risks.

Objectives and structure of this paper

This paper discusses the promises and perils of AI in health, and the key policy questions that policy makers will need to address in an uncertain landscape. The goal is to foster a shared understanding and to inform a G20 dialogue on AI on the main policy issues of AI use in the health sector. There is a need for better understanding of the most appropriate role of government in relation to AI applications in the health sector, so as to effectively manage the risks while not unnecessarily limiting the opportunities. This document showcases a number of applications of AI in health care, highlights some of the challenges associated with the design, use and implementation of AI in health, and considers the pertinence of the G20 AI Principles in the context of health care.

This report includes a collection of practical examples and avenues for discussion in the 2nd Digital Economy Task Force (DETF) meeting. It is not meant to be exhaustive and countries will certainly be in a position to share more national examples and challenges, potentially even first policy interventions. How AI changes health care, for better or for worse, will be in large part influenced by public policy, and AI is naturally a priority on national and international agendas including in the G20. Going forward, priorities for action should be established, including considering how to promote better health data and its governance, and to operationalise the G20 AI Principles in the health sector.

Questions for discussion

1. What are the challenges your government faces in operationalising the G20 AI Principles in the health sector?
2. What steps do you believe governments can take to ensure that AI in health is trustworthy?
3. Can you provide examples of actions and initiatives that you are currently undertaking to improve trustworthy AI in health?
4. Taking COVID-19 as a case study, how is your government using or planning to use AI as part of potential solutions, while ensuring trustworthy AI in such a context?

Table of contents

Figures

Boxes

1 Introduction

1. Much hype and great expectations currently surround Artificial Intelligence[1] (AI) to solve social and economic problems. AI is talked about as a central driver of the 4th industrial revolution, with an impact on humanity and on the planet as profound as those driven by the steam engine and electricity. AI holds great potential to enable the achievement of policy objectives – inclusive growth, productivity and the achievement of the Sustainable Development Goals (SDGs). Yet, the risks of harm are equally high. Like any powerful technology, AI can serve to entrench current social, economic and geopolitical problems, perhaps unintentionally. At worst, it can be deployed towards nefarious and destructive purposes.

2. The potential for AI in health is profound. From clinical uses in diagnostics and treatment, to biomedical research and drug discovery, to "back-office" and administration, it would seem there is almost no facet of health care provision and management where AI cannot be applied. The number of potential applications is growing every day. With health care systems facing significant longer-term challenges – including populations that are rapidly ageing and have multiple chronic diseases, gaps in workforce supply and skills, and growing health spending, a share of which is wasted or even harmful – and new threats like COVID-19, AI could make a huge difference in the coming years, and even weeks. The potential to tackle unwarranted variation in care, reduce avoidable medical errors, inequalities in health and health care, and inefficiencies and waste is particularly promising (OECD, 2017[1]). AI could have significant benefits not only under business-as-usual, but could also lead to better resilience and emergency preparedness, making health systems and societies more capable of responding to disease outbreaks like COVID-19.

3. At the same time, AI is also fuelling legitimate concerns and anxieties.

- The use of AI in everyday health care delivery and administration is still very limited, given there are serious difficulties in scaling up projects, and questions about the quality of health data.
- There are questions concerning the robustness of algorithms in the real world, and a policy and regulatory vacuum that greatly limits institutional and human capacity to realise the potential of AI.
- AI can be used to highlight existing inequalities and unwarranted variation in care, although there is a danger that, without proper control, AI could codify and reinforce biases and exacerbate inequality.

4. No single country or actor has all the answers to these challenges, and international co-operation and multistakeholder discussion are crucial to develop responses to guide the development and use of trustworthy AI for the wider good (OECD, 2019[2]).

[1] AI is a general-purpose technology that has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges (OECD, 2019[2]).

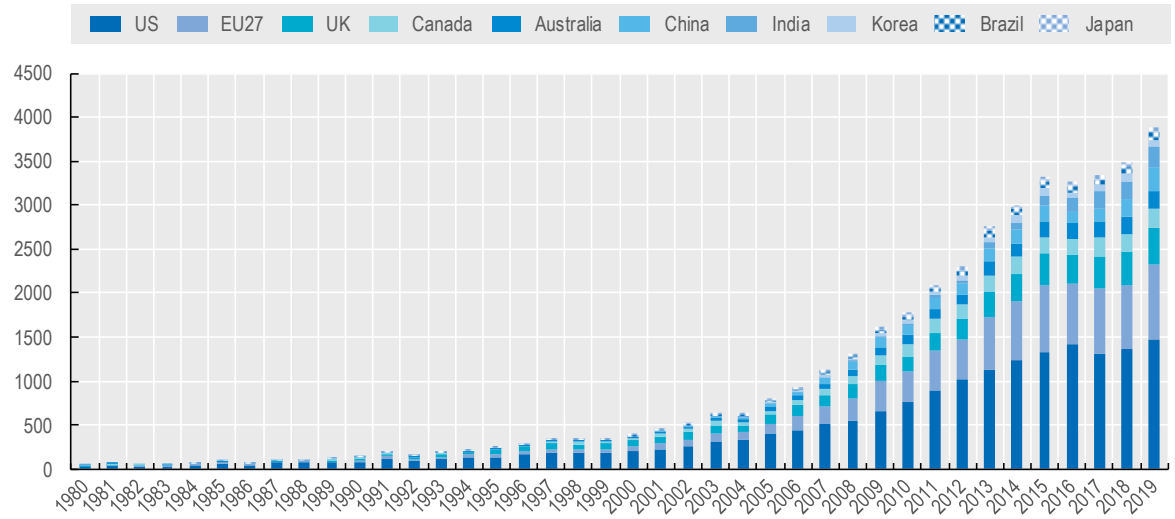
2 Artificial intelligence in health: profound potential but real risks

2.1. AI has profound potential to transform health care for the better

5. The use of AI in health in everyday practice is still extremely limited, yet the number of potential applications is growing by the minute. Applications range from the clinical settings, to biomedical research, to health system administration and management. Virtually every aspect of health care delivery seems amenable to AI use and implementation. The number of scientific publications relevant to AI in health has grown from just 36 in 1980 to close to 3 900 in 2019 (see Figure 2.1).

Figure 2.1. Scientific research on AI in health is booming

Number of relevant scientific publications in health, by country, from 1980 to 2019



Note: Please see methodological note (https://www.oecd.ai/assets/files/Methodology_20200219.pdf) for more information.
Source: OECD.AI (2020), visualisations powered by JSI using data from MAG, accessed on 3/3/2020, www.oecd.ai

2.1.1. AI promises to change clinical practice in the not so distant future

6. Much promise and research activity concerns the potential application of AI in the clinical setting, such as the automation of diagnostic processes, clinical decision-making and a range of other clinical

applications. A lot of activity to date has focused on **diagnostic imaging** (Neri et al., 2019[3]). There are many impressive examples[2] in the research setting where AI tools can perform as well as – or even better than – average clinicians, ranging from retinal scans to tumour detection in radiology. Another application of AI that has had success in the clinical research setting is **radiomics**: the extraction of certain features from multiple diagnostic images of one patient to produce a quantitative 'picture' of an anatomical region of interest. Such features may be used to *predict prognosis and response to treatment*. By building patients' radiological 'signatures', AI can enable accurate analysis and correlation between radiomics and other data (e.g. genomics, biopsies) that humans cannot[3].

7. In the surgical field, AI can aggregate diverse sources of information – including patient risk factors, anatomic information, disease natural history, patient values and cost – to predict the consequences of surgical decisions. AI-powered **remote-controlled robotic surgery** can improve the safety of interventions where clinicians are exposed to high doses of ionizing radiation, and makes surgery possible in anatomic locations not otherwise reachable by human hands (Svoboda, 2019[7]). As autonomous robotic surgery improves, it is likely that, in some cases, surgeons will oversee the movements of robots – and it is worth noting that, in most cases, currently robotic surgery does not result in superior outcomes than human surgeons.

8. In many medical interventions, customising clinical diagnostic reference levels based on appropriateness criteria and on patient characteristics – such as age, body mass index, vital signs and prior exposure to disease or risk factors – is an important risk management process. AI can be an optimising tool for assisting clinicians in providing a **personalised patient protocol**, in tracking the patient's dose parameters, and in providing an estimate of the risks associated with cumulative doses.

9. AI tools can also affect the daily workflow of a health care practice by assigning priority based on appropriateness criteria. **Clinical decision support systems** can assist referring physicians to choose the most appropriate investigative procedure based on the level of evidence for appropriateness, and the level of emergency. Such a system can be enhanced by AI to improve decision making speed and accuracy, optimising clinical workflow. For example, researchers at Clalit Research Institute, the largest health payer-provider network in Israel, have developed a model, based on a combination of regression and machine learning, with superior discriminatory power to that of previously reported models, that can be used to identify high risk of readmission early in the admission process (Shadmi et al., 2015[8]). In another example, in Kwa-Zulu Natal, South Africa, two doctors developed a cloud-based data analysis platform and recommendation engine that helps a very limited staff prioritise HIV patients based on their clinical risk, targeting potentially life-saving care to those who most need it (Combs, 2019[9]). The system also estimates the likelihood that patients will follow treatment, so that staff can follow up with patients at risk of non-compliance.

10. The process of reviewing the records of patients with unexpected outcomes in order to identify recommendations for improvement, known as a **clinical audit**, is an important but labour-intensive task for health professionals in all settings, and therefore a prime target for automation. A recent study compared human performance in generating audits on a neurosurgical ward, with that of an AI algorithm (Brzezicki et al., 2020[10]). The audits conducted by AI were significantly more accurate and logically

[2] Some of the numerous recent experimental applications of AI include determining skeletal age using pediatric hand radiographs (Halabi et al., 2019[4]), breast cancer detection in mammography and Magnetic Resonance Imaging (MRI), chest radiography interpretation, liver lesion characterisation on ultrasound and Computed Tomography (CT), brain tumour, and prostate cancer detection. For example, in Brazil, the Hospital do Câncer de Barretos in São Paulo, is using AI to more accurately assess the need for treatment of thyroid and prostate cancers (Izique, 2018[5]).

[3] For example, NITI Aayog and the Tata Memorial Centre Imaging Biobank, in India, use AI to generate imaging biomarkers for use in research studies, and support biological validation of existing and novel imaging biomarkers, hoping to eventually improve decision support in cancer treatment at low cost (Mahajan et al., 2019[6]).

consistent, for a fraction of the incremental resources compared to a human audit. For example, the mean time to deliver a report was 5.80 seconds for AI compared to 10.21 days for humans.

2.1.2. AI is already making a mark in biomedical research and public health

11. Biomedical and population health research seems to be more advanced compared to clinical applications of AI. This could be due to the lower barriers to adoption, the fact that the research setting is the scale of interest, as well as greater availability of capital and other resources. The exciting potential of combining AI with large datasets was demonstrated recently when Canadian company Blue Dot, which scours global media for information on a range of **infectious diseases**, spotted information about a "pneumonia of unknown cause" with 27 casualties in the People's Republic of China (hereafter "China"). It immediately issued an alert to its clients about potential risks of travelling to cities like Shanghai, Tokyo and Hong Kong, China. This occurred a week before the United States Centers for Disease Control and Prevention (CDC) and the World Health Organisation (WHO) issued alerts regarding the novel Coronavirus (now known as Coronavirus Disease or COVID-19).

12. Indeed the biomedical research setting is proving fertile ground for AI, where it might exponentially increase the probability for **new drug discovery** including important drugs such as novel antibiotics and antivirals. Researchers at the Massachusetts Institute of Technology (MIT) recently trained a deep learning algorithm to predict molecules' potential antimicrobial activity (Stokes et al., 2020[12]). The algorithm screened over one billion molecules and virtually tested over 107 million, identifying eight antibacterial compounds that were structurally distant from known antibiotics, with one effectively treated resistant infections in mice. In similar work, a deep learning model of small-molecule drugs was used to identify key biological mechanisms implicated in fibrosis and other diseases (Zhavoronkov et al., 2019[13]). As health systems are racing to find **treatments for COVID-19**, a machine learning model developed in London has discovered that a drug used in rheumatoid arthritis may be effective against the virus, and a Hong Kong, China-based company reported that its artificial intelligence algorithms have designed six new molecules that could halt viral replication (Mccall, 2020[10]). While drugs identified using AI will still need to be transitioned from the 'bench to the bedside', these examples highlight the exciting potential value of AI to discover urgently needed compounds such as antimicrobials (including antivirals).

13. Precision medicine is an area where AI fed by large volumes of personal health data can produce information that can help tailor medical treatment to the individual characteristics of each patient (Eichler et al., 2018[14]). Most medical treatments in use today work for a large majority of patients, but for some patients treatment either fails to deliver benefits or may even put them at risk of adverse events. Being able to pre-emptively identify those patients who are most likely to benefit, or otherwise, from therapy can lead to better health outcomes and more patient-centred care. Developing **precision medicine** turns the traditional research paradigm on its head – statistical noise and variation are the variables of interest, which cannot be feasibly achieved by prospective, traditional clinical trials. AI models based on large and varied personal health data are a key opportunity to make precision medicine a reality. To truly impact routine care, however, the data needs to represent the diversity of patient populations.

14. Disease **prediction and prevention** are another promising area for AI. Among other applications[4], researchers have demonstrated the ability of an algorithm to accurately predict the risk of emergency admission based on an individual's electronic health record data (Rahimian et al., 2018[15]). Coupled with data outside the health system (e.g. internet searches), such algorithms could be even more powerful,

[4] Machine learning algorithms using internet search and social media data have also been used by the City of Chicago, in the United States, to predict and pinpoint the source of foodborne disease outbreaks much faster than traditional inspection methods (OECD, 2019[15]). Researchers at the National University of Córdoba, in Argentina, have used neural networks to predict the likelihood of the emergence of an Aedes aegypti mosquito colony, a vector of viral diseases such as Dengue, Chicungunya and Zika, based on satellite images (Scavuzzo et al., 2017[16]).

although beset by privacy challenges that must be tackled at policy level. AI can also improve matching individuals to **clinical trials** (Lee and Lee, 2020[17]). Patients can be identified to enrol in trials based on more sources than clinical or administrative data. The criteria for including patients in a trial could take significantly more factors (genetic information) into account to target specific populations. This can enable trials to be smaller, shorter, set up more effectively and be therefore less expensive, without sacrificing statistical power. It can also address the documented problem of underrepresentation of minorities in clinical trials.

2.1.3. The application of AI in the 'back office' could have the most immediate impact

15. Health systems are notoriously complex. Providing services to individuals and populations involves a wide range of actors and institutions: patients, professionals, health care facilities and organisations, laboratories, imaging facilities, pharmacies, administrators, payers, and regulators. In parallel to the care provided, administrative workflow includes **scheduling, billing, coding and payment**. One of the principal and immediate applications of AI is to perform these mundane, repetitive tasks in a more efficient, accurate and unbiased fashion (NAM, 2019[18]). The back-office (e.g., scheduling, billing, coding and payment) is also a relatively safe testing ground for the technology's intended and unintended consequences, as errors tend to mostly carry administrative or financial risk, and only in select cases could they jeopardise patient safety.

16. Among the more obvious administrative uses of AI-based automation is **coding**: the process of extracting information from clinical records and codifying it using classifications such as the International Classification of Diseases (ICD) or Diagnosis-related groupings (DRGs). Coding is a complex, labour-intensive process, and coding accuracy is very important for reimbursement, administration and research. While computer-assisted coding has existed for more than a decade, AI can enhance the accuracy and transparency of clinical coding. AI tools are now available[5] that aggregate long, narrative clinical records with other relevant data (medication orders, imaging and laboratory tests) for deep learning and other AI models. Human review will still be required, but the processing 'grunt work' could be done by machines. In the short-term, AI may help human coders and create checks for policy makers and payers. In the long-term, near-full automation might be achieved but will undoubtedly rely on data quality and comprehensiveness, algorithm transparency and accuracy, and the trust of those relying on the outputs.

17. The ability of AI to analyse free text can be particularly powerful in a workflow where administrative decisions are based on narrative data, such as **prior authorisation** (PA), which is needed in most health systems (to some extent) to supply of health services and products to patients (Accenture, 2018[20]). PA requires the submission of patient information along with the proposed request and justification. Determination requires professional skill, analysis, and judgment. Automating this can improve the speed, consistency, and quality of decisions. With good governance, such a process could lead to fewer appeals and limit liability. There is a range of options to use AI in the context of PA. For example, AI methods could be used to triage cases to the appropriate level of reviewer. A more complex algorithm could find relevant information across one or more datasets to, for example, not only determine the eligibility of patient for a procedure, but also estimate the costs, the benefits and the risks associated with it. Moreover, lessons from applying this process at scale could be particularly useful in deploying similar AI functions in the clinical setting, such as triaging images for human review or automated reading, for example.

[5] For example, in China, researchers developed a deep-learning based natural language processing system to comb through more than 1.3 million electronic health records to extract relevant information and generate diagnoses (Liang et al., 2019[19]).

18. Much like AI can be taught to spot irregularities on medical images, algorithms can also learn to look for **fraudulent activity** in health care[6]. A variety of previously curated claims data (i.e. identified as fraudulent, valid, or requiring investigation) can be used to teach an algorithm to identify potentially fraudulent claims, including subtle systematic cases of over-servicing and upcoding (i.e. using a code for a more expensive medical service than was performed), as well as potential underuse of available services. Combining claims with other data (e.g. clinical and demographic) can increase the accuracy of the audit process. For example, in addition to age and co-morbidities, information on a patient's entire medical history may help ascertain if a procedure was in fact necessary, and if additional payment adjustments for case complexity are justified, or if a potentially beneficial treatment may have been missed.

19. **Scheduling** is another example where AI can add value, and almost immediately. A no-show is not only an inconvenience but also a waste of highly qualified human resources who are typically in high demand. Algorithms fed on historical data can predict which patients may not attend and take proactive action to manage this. Beyond blanket or even targeted reminders, AI can address a patient's needs and queries[7]. These uses may seem trivial, but considering that around one fifth of all health spending is estimated to be wasted on inefficiencies, including fraud, a systematic reduction represents hundreds of billions of dollars each year that could be diverted towards actual care (OECD, 2017[1]).

2.2. Applications of AI in health raise legitimate concerns and anxiety

20. While there is undoubtedly enormous potential for AI to transform almost every aspect of the health sector, the use of AI in everyday health care is still very limited. There are serious difficulties in scaling up projects to the level of health systems, due to, among other things, questions concerning the robustness of algorithms in the real world, a lack of high quality health data, and a policy and regulatory vacuum that greatly limits institutional and human capacity to realise the potential of AI.

2.2.1. AI in health is not yet robust: for every success story there is a cautionary tale

21. Unfortunately, AI applications in health have been beset by misfires and setbacks, with hype often clashing with reality. A recent study reviewed dozens of studies claiming an AI performs better than radiologists, finding that only a handful were tested in populations that were different from the population used to develop the algorithms (Reardon, 2019[21]). The difficulty to scale certain AI applications is often due to trivial factors. For example, the way different facilities label their images can confuse an algorithm and prevent the model from functioning well in another institution with a different labelling system. This serves to highlight that most AI in health is actually **artificial narrow intelligence**, or "weak" AI, designed to accomplish a very specific problem-solving or reasoning task, and unable to generalise outside the boundaries within which the model was trained (see Box 2.1).

22. The majority of AI applications in health rely on machine learning methods, ranging from linear and logistic regressions, decision trees and principle component analysis to deep neural networks. They usually rely on large amounts of training data to make predictions. Because these **methods are narrowly focussed** on a specific task and trained using a specific set of data, these algorithms may not work well when given input data that is even just slightly different from the training data. This is why something as

[6] For example, SAS is working with DentaQuest, a health insurer in the United States with 24 million members, to detect fraud using SAS Enterprise Miner, a data mining and analytics platform that uses AI.

[7] For example, Northwell Health – a US health network – has launched a text-based patient outreach programme aimed at reducing no-show rates for colonoscopy procedures. The colonoscopy health chat platform is a flexible algorithm that aims to give patients the information they need so that they feel comfortable and motivated to get screened. The chatbot addresses questions and concerns related to the procedure, its benefits, and reminds the patient of the date and location as the procedure draws closer.

simple as a difference in labelling can cause AI models trained in one setting to malfunction in another, and why such models have not really scaled much across health systems.

23. Another consequence of this **heavy dependence on the input data** is that models that have been trained on a certain patient population may not function well when fed with data for a different patient population. Most AI applications in health are still in research and development stages, concentrated in a few countries and regions as illustrated in Figure 2.1 by the fact that the top five countries/regions (US, EU27, UK, Canada and Australia) have more than six times more research papers cumulatively than the bottom five. As such, most of the data used to train these models is from Western, educated, industrialized, rich, and democratic (i.e., "WEIRD") populations. It is almost certain that algorithms used to explain or predict human behaviours based mainly on care patterns for one population will be biased (NAM, 2019[18]). For example, an AI algorithm used to identify patients with complex needs in the United States has been shown to suffer from racial bias, assigning lower risk to Black patients compared to White patients. Using health costs as a proxy for health needs, the algorithm learned that since less money is spent on Black patients who have the same level of need, Black patients are healthier than equally sick White patients (Obermeyer et al., 2019[22]).

Box 2.1. Artificial narrow intelligence and artificial general intelligence

Most AI in health is actually artificial narrow intelligence, incapable of versatile abstract learning

Artificial narrow intelligence or "weak" AI is the current state-of-the-art. The most advanced AI systems available today, such as Google's AlphaGo, are still "narrow". To some extent, they can generalise pattern recognition such as by transferring knowledge learned in the area of image recognition into speech recognition. However, the human mind is far more versatile. Applied AI can be contrasted to a (hypothetical) artificial general intelligence, in which autonomous machines would become capable of general intelligent action. Like humans, they would generalise and abstract learning across different cognitive functions. Artificial general intelligences would have a strong associative memory and be capable of judgment and decision-making, solving multifaceted problems, learning through reading or experience, creating concepts, perceiving the world and itself, inventing and being creative, reacting to the unexpected in complex environments and anticipating.

Source: (OECD, 2019[2])

24. A related challenge is **overfitting**, which occurs when an AI model learns statistical irregularities specific to the data on which it is trained. Unless the training data are vast (and therefore difficult and costly to create) the model may confuse irrelevant noise with the actual signal. An overfitted model will not generalise to different input data. Overfitting was one of the problems in the IBM Watson cancer initiative, where the model was trained on hypothetical data and then graduated to real clinical situations too quickly (Strickland, 2019[22]).

25. The large majority of machine-learning-based prediction models are based on **correlation, not causation** (NAM, 2019[18]). Previous studies have identified counterintuitive associations that lead to nonsensical predictions. For example, a model that predicts risk of death for a hospitalized individual with pneumoniae learned that patients who have asthma and pneumoniae are less likely to die than patients who only have asthma, because patients with asthma and pneumoniae receive more aggressive treatment and thus have lower mortality rates. In another example, the time a lab value is measured can often be more predictive than the value itself (e.g. if it is measured at 2am).

26. Algorithms that learn from human decisions will also **learn human mistakes, biases and stereotypes**. Yet, the AI sector is extremely gender and race imbalanced (AI Now Institute, 2019[23]),

suggesting that biased and stereotypical predictions might not be flagged by developers working to validate model outputs. For example, Apple's HealthKit, an application to track intake of selenium and copper, neglected to include a women's menstrual cycle tracker until iOS 9; the development team reportedly did not include any women (NAM, 2019[18]).

27. Finally, the predictions of an AI model must eventually be operationalised in the form of information systems: e.g. an alert or pop-up window within electronic health record software. There is a body of evidence showing that the implementation of health information systems can result in unintended consequences. These include alert fatigue, imposition of additional workloads for clinicians, disruption of interpersonal (including doctor-to-patient) communication styles, and generation of specific hazards that require a higher level of vigilance to detect. Growing numbers of health workers are already stretched (Dyrbye et al., 2017[24]), with some suffering from **change fatigue**: getting tired of new initiatives and the way they are implemented (Garside, 2004[25]). Against this backdrop, the black-box nature of AI algorithms may result in either resistance from clinicians to adopt and use their predictions, or a blanket acceptance of their outputs with little critical assessment of the potential for biased and suboptimal predictions. The interface between the algorithmic world and the brick-and-mortar world is key to success.

2.2.2. Poor health data governance means AI requires a lot of human curation

28. For most methods and applications, AI typically needs large amounts of data to train and prepare for real-world application. If these data even exist – which is not guaranteed – it is likely they will need human **curation**, including, for example, the need to be stratified by patient cohort, segmented to extract the region of interest for AI interpretation, filtered, cleaned and labelled. This process can be very time- and labour-intensive, and is not easy – if even possible in the near future – to automate.

29. Curation is essential to ensure a robust validation of 'ground truth', a fundamental concept in most types of AI. This means validating the output as a true positive or negative when the machine is in the learning phase. This can be problematic in health, where **data are notoriously messy**, and classifications or interpretations of the underlying data can be wrong to begin with. For example, in the clinical setting, the presence or absence of pathology is often not binary but a matter of opinion. In oncology, there can be disagreement on what constitutes cancer requiring medical intervention. Experienced pathologists will often disagree about the histopathology and diagnosis, particularly in early-stage lesions. If the aim of AI is to help detect cancer early, this **disagreement presents a conundrum** on how to train the algorithm, as the resulting AI tool should not over- or under-diagnose tumours.

30. Even if there were perfect agreement on what the 'ground truth' is, most existing medical data is not readily available for use in AI algorithm development, or of high enough quality (not easy to exchange, process or interpret, riddled with errors). While most people think health care is awash with big data, in many countries the reality is much closer to "**a large number of disconnected small data**" (Lehne et al., 2019[26]). A central source of patient-level data for AI development – the electronic health record – is emerging as a key resource for dataset development and research in only a handful of OECD countries with standardised, interoperable electronic health record (e-HR) systems offering "one patient, one record" national data (Oderkirk, 2017[27]). Against this backdrop of poor data governance and data quality, it is unlikely that AI algorithms can be used in the real world without extensive human input and investment to improve the quality of routinely collected clinical data.

2.2.3. The carbon footprint of AI and its relation to the burden of disease are still unclear

31. An **area of great uncertainty** surrounding AI in general is its carbon footprint. Training a single AI model can emit as much carbon as five cars in their lifetimes, a figure that could be but a minimum estimate as training a single model is only the beginning of an algorithm's lifecycle (Hao, 2019[28]; Strubell, Ganesh and McCallum, 2019[29]). While a growing appetite for digital services could mean the data centres that

power them emit a rising share of global greenhouse gases, a shift to cloud computing has led to massive efficiency gains in recent years (Masanet et al., 2020[30]). AI itself could be used to promote a circular economy and increase energy efficiency, as illustrated in Google DeepMind's use of machine learning in Google's data centres to reduce energy use for cooling by 40% (DeepMind, 2016[31]). With one estimate indicating health systems in the OECD, China and India already account for 5% of total emissions, more than aviation or shipping (Pichler et al., 2019[32]), **AI could have both positive and negative net effects**.

32. If AI in health turns out to be a net contributor to greenhouse gas emissions, this would mean that an activity within the health sector would itself be associated with an increase in the burden of disease. Moreover, given the high entry costs to developing AI models, it is likely that development will take place mostly in high-income settings, with the majority of the climate effects on health being felt in low-income settings. On the other hand, if AI leads to more energy-efficient health care systems, then the impact on vulnerable areas could be disproportionately beneficial. As uncertainty remains today, **energy use associated with AI in health should be monitored and studied**.

3 Priority for policy: beware the hype and start with real problems

33. AI has the potential to help achieve agreed and important policy objectives such as inclusive growth and the Sustainable Development Goals. It can also amplify and entrench current social and economic and geopolitical problems. At worst, it can be deployed towards nefarious and destructive purposes. Which path the world will take with regard to AI is principally a policy choice. While interest in AI in health is already growing and the private sector is moving fast, widespread use remains limited. This provides an opening for policy makers to get ahead of the curve, and discuss how best to capitalise on the real opportunities that AI affords, while considering mechanisms to ensure risks are prevented, minimised and contained. In the absence of a well-defined ready-to-use menu of policy options, countries should promote a multilateral inclusive discussion on a plan of action that promotes trustworthy AI in health.

34. The G20 AI Principles provide a framework to guide the discussion. Their value-based principles aim to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values. They identify five complementary value-based principles:

- inclusive growth, sustainable development and well-being;
- human-centred values and fairness;
- transparency and explainability;
- robustness, security and safety;
- and accountability.

35. In addition to and consistent with these value-based principles, the G20 AI Principles note five recommendations to policy-makers on national policies and international co-operation for trustworthy AI:

- investing in AI research and development;
- fostering a digital ecosystem for AI;
- shaping an enabling policy environment for AI;
- building human capacity and preparing for labour market transformation;
- and international co-operation for trustworthy AI.

36. The principles and recommendations lend themselves well to the health setting. In fact, there are synergies between them and specific instruments related to the health sector (e.g. the OECD Council Recommendation on Health data Governance, which places emphasis on transparency, trust and communication as well as the need for international cooperation to improve health data standards and interoperability). Of particular relevance to health are perhaps the following set of principles and recommendations: fostering a digital ecosystem for AI by supporting safe, fair, legal and ethical data sharing; operationalising the AI principles, namely ensuring transparency, explainability and accountability of AI outputs; implementing regulatory oversight and guidance that encourages innovation in trustworthy AI; building human capacity (among health workers but also patients and populations who must be

comfortable with 'intelligent' machines); and allocating long-term public investment (in a sector where resources are increasingly scarce).

3.1. Fostering a digital ecosystem for AI through health data governance for safe, fair, legal and ethical data sharing

37. AI cannot function as intended without good data from which to learn. High-quality, representative and real-world data are essential to minimise the risk of error and bias. Creating an environment where such data – especially personal health data – are available to AI researchers and developers in a secure way that respects individuals' privacy and autonomy is fundamental. This requires frameworks for strong health data governance, within and across countries.

38. In line with the G20 AI Principles, frameworks for health data governance should emphasise transparency, public communication and stakeholder engagement, explicitly highlighting **the importance of trust** (OECD, 2016[33]). Lack of trust among patients, the public, data custodians and other stakeholders, in how data are used and protected is a major impediment to data use and sharing. Personal health data are very sensitive, and privacy is understandably one of the most frequently cited barriers to using them. Yet, the potential benefits of using personal health data to generate new knowledge cannot be minimised, for example in the context of testing much needed drugs and vaccines (as currently highlighted by the COVID-19 crisis). Health care leaders should work to advertise the benefits of using health data, changing the discourse that sees use of data as the only risk and that ignores the foregone benefits to individuals and societies of failing to put data to work (OECD, 2019[15]). It is also essential to dispel the idea that there is a trade-off between data protection and secondary use of health data. A risk management approach and careful implementation of good practices can enable both data protection and its use. Updated periodically, formal risk management processes could include unwanted data erasure, re-identification, breaches or other misuses, in particular when establishing new programmes or introducing novel practices.

39. For a number of reasons (e.g. representativeness and breadth of input data), many applications of AI in health would gain considerably from **cross-border collaboration** in the processing of personal health data for purposes that serve the public interest. This includes identifying and removing barriers to effective cross-border collaboration in the processing of personal health data, as well as engaging with relevant experts and organisations to develop mechanisms to enable the efficient exchange and interoperability of health data, including by setting standards for data exchange and terminology (OECD, 2016[33]). Sharing data across jurisdictions is central to advance AI research in areas such as cancer and rare diseases, as it requires sufficiently large, representative and complete data (and could potentially reduce the carbon footprint associated with AI). Cross-border data sharing is also crucial during pandemics (e.g. COVID-19), when infection spreads globally and concerted action is needed.

40. The latest evidence suggests that countries are lagging in their implementation of robust, consistent governance frameworks for the use of personal health data (OECD, 2019[15]). Given the fundamental importance of good data in AI, failure to implement strong governance models will hinder and ultimately stall the potential benefits of this powerful technology. Given the **need for global coordination** in this regard, many AI applications in health will only be as good as the weakest link. This underscores the critical relevance also of the G20 AI Principles' recommendation for international co-operation for trustworthy AI.

41. Harmonisation and interoperability of the laws and policies governing health data enables cross-country collaboration. OECD countries are divided regarding legal permission to share data across borders for research and statistical uses in the public interest, even if the data are de-identified first. Among 18 high-income countries, only seven had laws and policies that could permit the majority of national health datasets to be shared with a foreign researcher working in the non-profit or governmental sectors for a project within the public interest (OECD, 2019[15]). The European Union's General Data Protection

Regulation (GDPR) fosters improvement in this regard among EU countries. The GDPR is a central feature of the Union's ambition to make health data more structured, interoperable, portable across borders, secure and respectful of individual's privacy. These laws and policies have the potential to advance the availability of high-quality, representative data for the development of AI models. Importantly, the GDPR puts health data in a special category that can be used for secondary purposes such as research deemed to be in the public interest and sets out the conditions for approval of cross-border data sharing. The GDPR has the potential to influence legal reforms outside the EU, particularly among countries seeking research partnerships and data exchanges with EU countries, and thus it has the potential to promote harmonisation in laws and policies regarding cross-border sharing beyond the EU.

3.2. Operationalising the values-based G20 AI Principles

42. Agreement on the values-based elements of the G20 AI Principles was an important achievement but represents only the start. Operationalising them consistently across countries will be the challenge. For example, AI actors should commit to **transparency and explainabiility** (a G20 AI Principle) and responsible disclosure regarding AI systems. Of particularly relevance to health care is the principle that those affected by an AI system – adversely or otherwise – should be aware that AI was used, and be able to understand the outcome and potentially challenge it, based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision (OECD, 2019[35]). Implementing this **in practice can be technically challenging**, as illustrated in the following excerpt from a recent report of the United States National Academy of Medicine (NAM, 2019[18]):

> *Many contemporary AI systems are deployed on cloud-based, geographically distributed, nondeterministically parallelized, spatially arbitrary computing architectures that, at any moment, are physically unidentifiable. To create and maintain a full log of each processor that contributed in some part to the execution of a multielement ensemble model AI is possible in principle but would likely be cost-prohibitive and too cumbersome to be practical. Therefore, the limited traceability and fundamental non-recreatability and non-retestability of a patient's or clinician's specific execution of an AI system that may have contained a fault or that produced errors or failures—untoward, unexpected deviations from its specifications, validation testing, and hazard analysis—may pose particular problems for regulators, courts, developers, and the public.*

43. AI actors should also be accountable for the proper functioning of their algorithms, within the scope of their own roles. Legal clarification regarding **accountability** (a G20 AI Principle) and responsibility for AI model outputs is important. The European Union recently published a report stating that manufacturers of products or digital content incorporating emerging digital technology should be **liable for damage** caused by defects in their products, even if the defect was caused by changes made to the product under the producer's control after it had been placed on the market (European Union, 2019[36]). Similarly, in 2018, the United States Food and Drug Administration (FDA) approved the first algorithm that can make a medical decision without the need for a physician to look at a diagnostic image (Reardon, 2019[21]). Because no doctor is involved, the company that developed the algorithm has assumed legal liability for any medical errors.

44. Ensuring the **robustness, security and safety** (a G20 AI Principle) of AI algorithms and applications is paramount, and the FDA has recently proposed a set of guidelines to manage algorithms that evolve over time. Among them is an expectation that developers **monitor how their algorithms are changing** to ensure they continue to work as designed and asking them to notify the agency if they see unexpected changes that might require re-assessment (Reardon, 2019[21]).

45. Operationalising the values-based AI Principles will require considerable investment of financial and political capital. For example, AI in health will be a lucrative business, and it will take a lot of political will to enact legislation that ensures openness, transparency and accountability. Even ideological differences can present obstacles to guaranteeing AI in health is guided by **human-centred values and**

fairness (a G20 AI Principle). For example, some advocate that personal health data are like any commodity owned by the data subject who should have the freedom to sell or exchange them. While the question of ownership can be debated, there is little doubt that such commodification of health data will incentivise poorer, more disadvantaged people to sell theirs. Ethical underpinnings of such a policy position aside, purely from a technical standpoint this will create sample bias in the data used to train AI models. This could increase representation of patients of lower socio-economic status in AI algorithms, but there are other ways to increase representation that do not involve having a group of patients sell their medical data.

46. With respect to **inclusive growth, sustainable development and well-being** (a G20 AI Principle), because developing and using AI requires substantial resources and investments, and because low-resource settings might not be as ready to act on AI predictions as well-resourced settings, there is a risk that existing inequalities in health between high-income settings and low-income settings (within and across countries) could persist, or even be amplified, for example as AI algorithms developed in high-income settings are either not used in low-income settings or used with poor outcomes. AI technology developed in environments that exclude patients of different socioeconomic, cultural, and ethnic backgrounds, can lead to poorer performance in the groups that potentially most stand to benefit. Decision makers must address these challenges as they build policy and regulatory frameworks, and invest in operational capacity.

3.3. Shaping an enabling policy environment for AI by putting in place regulation and guidance that promote trustworthy AI

47. AI is new territory for health policy makers, providers and the public. In terms of fostering trustworthy AI that delivers for patients and communities, an **enabling policy environment for AI** that includes a risk management approach is needed (in line with G20 AI Principle of robustness, security and safety). One way is through **regulatory sandboxes** – contained testing grounds for new approaches and business models. Regulatory sandboxes allow developers and health care providers to test and evaluate innovative products, services and business models in a live environment with the appropriate oversight and safeguards (ring-fencing wider health systems from risks and potential unintended consequences). The United Kingdom's Care Quality Commission and the Singaporean government are using regulatory sandboxes to test new (digital) health models.

48. Japan, for example, has developed AI Utilization Guidelines to enhance the outcomes of AI models, drawing on the G20 AI Principles (OECD, 2019[2]). The Guidelines also specify that AI actors should create and publish an AI usage policy and notify users, consumers, and others, so that they are aware of the use of AI. While Japan has traditionally been near the bottom of the pack in making personal health data available for secondary purposes, the government has recently set in motion legislative reforms to address this (OECD, 2019[15]).

49. It is encouraging to see wide agreement regarding the need for AI principles and values, with at least 84 public-private initiatives describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI (Jobin, Ienca and Vayena, 2019[34]). However, a multitude of frameworks poses a risk to international cooperation. The onus is on countries to draw on a set of value statements to develop and implement the necessary policies, **regulations and legal frameworks**. Consistency across jurisdictions will be to everybody's advantage, and the practical and technological challenges to several of the AI principles can perhaps be better overcome through **international co-operation for trustworthy AI**.

3.4. Building human capacity and preparing the workforce for a labour market transformation in health

50. To date, there is **no evidence to suggest that AI will replace humans in health care**, but it there is plenty to suggest that it will fundamentally augment human tasks, skills and responsibilities. Given the scale at which AI could change the healthcare landscape, the way health workers – and indeed the entire workforce – are educated, trained and socialised will need to adapt. The approach will need to be multidisciplinary, involving AI developers, implementers, health care system leadership, frontline clinical teams, ethicists, humanists, and patients and patient caregivers, as each provides their own point of view and lived experience – each one should be heard (NAM, 2019[18]).

51. As the National Academy of Medicine report points out (NAM, 2019[18]), **new jobs and professions** will be needed to realise the potential benefits of AI in health: trainers, explainers and sustainers. Trainers will provide meaning, purpose, and direction; explainers will use their knowledge in both the technical and application domains to explain how AI algorithms can be trusted to support decisions; and sustainers will help maintain, interpret, and monitor the behaviour and unintended consequences of AI systems.

52. A number of **countries are already preparing**. In France, a collaboration between Gustave Roussy, one of Europe's leading cancer hospitals, and two engineering schools in Paris, École des Ponts ParisTech and CentraleSupelec, hopes to train young computer scientists to understand medicine, and conversely, to train medical researchers to better understand the basics of artificial intelligence. In the United States, about a dozen fellowships are offered to train budding doctors in a range of engineering approaches, including artificial intelligence. Australia, Canada, Norway, Switzerland, New Zealand, and the UK have all completed reviews or established regular processes to assess how technological developments will change skill requirements, roles and functions of health care staff.

3.5. Investing strategically and sustainably in AI research and development

53. Preparing health systems to manage the risks and make the most of AI requires long-term strategic investment. Strategic, **coordinated and sustained resourcing is needed** to ensure that AI leads to desirable health, social and economic outcomes and takes a similar trajectory to successful industrial revolutions of the past. Public resources are and always will be scarce, but they need to be found given the profound opportunities for better and more equitable health outcomes brought by AI, and to act as a counter-weight to private investment (see Box 3.1).

54. Private capital is piling into the AI field (OECD, 2018[37]). **Private investment** doubled from 2016 to 2017, reaching USD 16 billion in 2017. AI start-ups attracted 12% of worldwide private equity investments in the first half of 2018. For example, in pharma, AI-based drug discovery start-ups raised more than $1 billion in funding in 2018. At least 20 separate partnerships have been reported between major pharma companies and AI-drug-discovery tech companies (Freedman, 2019[38]). Pfizer, GlaxoSmithKline and Novartis are among the pharma companies said to have also built substantial in-house AI expertise.

55. Besides potentially developing AI tools themselves, governments and public institutions should devote resources towards developing the guardrails that ensure this technology does maximum good and minimal harm – and the **checks and balances** to steer the private sector in the right direction. This includes establishing and maintaining sound policy frameworks, building institutional and human capacity to verify AI technology where needed, and use it in an appropriate and cost-effective manner.

Box 3.1. AI applications need to add value to health and policy

Balancing the prospective benefits and risks of private sector AI initiatives in health

In May 2019 the U.S. Food and Drug Administration approved Vyndaqel (tafamidis meglumine) and Vyndamax (tafamidis) for the treatment of heart disease caused by transthyretin mediated amyloidosis (ATTR-CM), a rare type of cardiomyopathy that can result in shortness of breath, fatigue, heart failure, loss of consciousness, abnormal heart rhythms and death. Efficacy of the drugs in treating ATTR-CM was demonstrated in a randomised clinical trial of 441 patients (FDA, 2019[39]). These are the first FDA-approved treatments for ATTR-CM, developed by FoldRx, a subsidiary of Pfizer.

Pfizer has also developed an AI algorithm to help identify patients at risk of ATTR-CM, using medical claims and electronic health record data. There is evidence that ATTR-CM is under-diagnosed in the early stages of the disease, so combining a validated AI-powered screening tool with an effective FDA-approved treatment could lead to a radical reduction in the burden of heart failure due to ATTR-CM. While Vyndaqel and Vyndamax have been reviewed and approved by the FDA, the AI algorithm used to screen patients has not. Given the black-box nature of AI algoritms, the limits to generalisability from one setting to another, the risk of false positive predictions resulting in over-diagnosis and unnecessary treatment (with potential side effects), and the financial interest to Pfizer in detecting cases, it is crucial that the AI algorithm be independently reviewed by a trusted organisation, to establish its accuracy, robustness and validity. Without such a review, it is unlikely that payers, providers and regulators will promote, adopt and use the algorithm widely, potentially delaying a much needed innovation.

Moreover, given the limited number of patients involved in the clinical trial of Vyndaqel and Vyndamax, it would be advisable to monitor the drugs' effectiveness and safety in real-world clinical settings. Observational studies could be combined with monitoring the performance of the AI-based diagnostic tool, and potentially compare it with other statistical models. A publicly funded and fully independent review and approval of the AI-powered screening tool would provide the evidence needed to make more accurate and earlier diagnosis a reality for patients, leading to better health outcomes, and more efficient and less wasteful health care.

Source: Pfizer AI algorithm from personal communication with Business at OECD (2020).

56. Gauging the economic benefits of AI, and its superiority over conventional and cheaper techniques must also be an important consideration given the comparative costliness of AI. A recent systematic review found that very few publications assess the **economic impact of AI**, with no studies having a methodologically complete cost impact analysis (Wolff et al., 2019[40]). Public investment is needed to measure the broader operational costs of AI infrastructure and services, and results must be compared to existing alternatives.

57. The prognosis is not favourable. National **health systems typically underinvest** in information systems given the paramount importance of information, communication and knowledge in this sector (OECD, 2019[15]). Plainly put, fiscal space to invest in guiding AI in health must be found. Fortunately, intelligent deployment of AI (and digital technology more broadly) provides opportunities to reduce waste and deliver efficiencies in an inefficient and wasteful industry. Unlike a drug or medical device, AI is a 'general purpose' technology that can be deployed in just about any facet or activity of the health industry. Rather than create new procedures or activities (although there are valuable research applications of AI - such in new drug discovery), AI can principally be used to make existing clinical and administrative processes more effective, efficient and equitable. With around one fifth of health spending being wasteful or even harmful, this is a major opportunity to improve outcomes and value. In the medium-run the investment may pay for itself.

4 Conclusion

58. AI technology is a pressing challenge for policy makers, not least because it is poorly understood. Erik Brynjolfsson – formerly Director of the MIT Initiative on the Digital Economy and the MIT Center for Digital Business and now with the Stanford Institute for Human-Centered Artificial Intelligence– describes the current attitude towards, and understanding of, AI by some industry and government leaders as somewhat naïve (AEI, 2020[41]). Many see it as magic, a welcome panacea to all problems. Like other hyped technologies past and present, the term is often used uncritically. **Expectations must be managed** as most AI in health does not currently even begin to replicate what most would consider 'intelligence' – including consciousness or creativity – and is not expected do so in the near future.

59. Most AI applications in health are just programmes designed to optimise a mathematical function. AI is created by humans and is therefore inherently prone to error. It is also imbued with the biases of the humans who programme them as well as biases present in the data on which it learns. Also, more involved AI models like deep neural networks can be only marginally superior to more conventional statistical methods like logistic regression (Tiwari et al., 2020[42]). AI technology should not be used like the proverbial hammer to crack a nut, and lead to misallocation of financial, computing and natural resources (if the climate impact is considered). These are key risks that policy makers and other actors must be aware of.

60. It **will likely take time, effort and investment** to realise the potential of AI, and success will depend not only on the merits of the algorithms, but largely on the redesign and reorganisation of health services and workforce skills (David, 1990[43]). Simply applying technology to an existing broken system will not work by itself nor lead to significant transformation or benefit. Electrification, for example, contributed immensely to global welfare, but success was not pre-ordained. It was the result of a great deal of planning, regulation, co-operation and public investment. Incentives and market mechanisms were certainly deployed, but kept firmly in check by strong policy frameworks based on agreed objectives, values and principles.

61. There is currently **a window of opportunity** to promote an inclusive, informed and balanced discussion at international level of the real opportunities and risks of AI in health. Such a discussion would benefit greatly from the contribution of different stakeholders and actors, within and beyond AI. While applications of AI in health remain fairly rare, the field is moving quickly, and time is running out for policy makers to be proactive and stay ahead of the curve. The G20 AI Principles provide a framework to guide the design of policy actions. In their efforts to operationalise the Principles, governments and policy makers can start by considering the following questions:

- What are the challenges in operationalising the G20 AI Principles in the health sector?
- What steps can governments can take to ensure that AI in health is trustworthy?
- Are there examples of actions and initiatives currently in place to improve trustworthy AI in health?
- Using COVID-19 as a case study, how are governments using or planning to use AI as part of potential solutions, while ensuring trustworthy AI in such a context?

Annex A. The National Academy of Medicine's recommendations for AI in health

62. The United States National Academy of Medicine recently published a special report summarising current knowledge on AI in health to offer a reference document for relevant health care stakeholders. It is a comprehensive and well-documented report, with sections on the current and near-term AI solutions; the challenges, limitations, and best practices for AI development, adoption, and maintenance; the legal and regulatory landscape for AI tools designed for health care application; and the need for equity, inclusion, and a human rights lens for this work. It also outlines key considerations for moving forward, including recommendations (NAM, 2019[19]):

- Beware of marketing hype, but recognize real opportunities.
- Seek out robust evaluations of model performance, utility, vulnerabilities, and bias.
- There should be a deliberate effort to identify, mitigate, and correct biases in AI tools.
- Demand transparency in data collection and algorithm evaluation processes.
- Develop AI systems with adversaries (bad actors) in mind.
- Prioritize education reform and workforce development.
- Identify synergy rather than replacement.
- Use AI systems to engage, rather than stifle, uniquely human abilities.
- Use automated systems to reach patients where existing health systems do not.

References

Accenture (2018), *The Intelligent Payer: A survival guide*, https://www.accenture.com/_acnmedia/pdf-82/accenture-intelligent-payer-survivor-guide.pdf. [15]

AEI (2020), *Erik Brynjolfsson: Can AI help us overcome the productivity paradox?*, https://www.aei.org/multimedia/erik-brynjolfsson-can-ai-help-us-overcome-the-productivity-paradox/ (accessed on 4 March 2020). [38]

Angus, D. (2020), *Randomized Clinical Trials of Artificial Intelligence*, American Medical Association, http://dx.doi.org/10.1001/jama.2020.1039. [42]

Brzezicki, M. et al. (2020), "Artificial intelligence outperforms human students in conducting neurosurgical audits", *Clinical Neurology and Neurosurgery*, Vol. 192, p. 105732, http://dx.doi.org/10.1016/j.clineuro.2020.105732. [7]

Combs, V. (2019), *South African clinics use artificial intelligence to expand HIV treatment - TechRepublic*, https://www.techrepublic.com/article/south-african-clinics-use-artificial-intelligence-to-expand-hiv-treatment/ (accessed on 10 March 2020). [6]

David, P. (1990), *The dynamo and the computer: An historical perspective on the modern productivity paradox*, American Economic Association, http://dx.doi.org/10.2307/2006600. [40]

DeepMind (2016), *DeepMind AI Reduces Google Data Centre Cooling Bill by 40%*, https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40 (accessed on 9 March 2020). [27]

Dyrbye, L. et al. (2017), "Burnout Among Health Care Professionals: A Call to Explore and Address This Underrecognized Threat to Safe, High-Quality Care", *NAM Perspectives*, Vol. 7/7, http://dx.doi.org/10.31478/201707b. [20]

Eichler, H. et al. (2018), "Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?", *Clinical Pharmacology and Therapeutics*, Vol. 105/4, pp. 912-922, http://dx.doi.org/10.1002/cpt.1226. [11]

European Union (2019), *Liability for Artificial Intelligence and other emerging digital technologies: Report from the Expert Group on Liability and New Technologies – New Technologies Formation*, http://dx.doi.org/10.2838/573689. [32]

FDA (2019), *FDA approves new treatments for heart disease caused by a serious rare disease, transthyretin mediated amyloidosis*, https://www.fda.gov/news-events/press-announcements/fda-approves-new-treatments-heart-disease-caused-serious-rare-disease-transthyretin-mediated (accessed on 4 March 2020). [36]

Freedman, D. (2019), *Hunting for New Drugs with AI*, Nature Research, http://dx.doi.org/10.1038/d41586-019-03846-0. [35]

Garside, P. (2004), "Are we suffering from change fatigue?", *Quality & safety in health care*, Vol. 13/2, pp. 89-90, http://dx.doi.org/10.1136/qshc.2003.009159. [21]

Halabi, S. et al. (2019), "The rSNA pediatric bone age machine learning challenge", *Radiology*, Vol. 290/3, pp. 498-503, http://dx.doi.org/10.1148/radiol.2018180736. [43]

Hao, K. (2019), *Training a single AI model can emit as much carbon as five cars in their lifetimes*, MIT Technology Review, https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/?utm_source=newsletters&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement (accessed on 4 March 2020). [24]

Izique, C. (2018), *Artificial intelligence at the service of cancer diagnosis*, http://pesquisaparainovacao.fapesp.br/artificial_intelligence_at_the_service_of_cancer_diagnosis/819 (accessed on 10 March 2020). [47]

Jobin, A., M. Ienca and E. Vayena (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, Vol. 1/9, pp. 389-399, http://dx.doi.org/10.1038/s42256-019-0088-2. [33]

Lee, C. and A. Lee (2020), "How Artificial Intelligence Can Transform Randomized Controlled Trials", *Translational Vision Science & Technology*, Vol. 9/2, p. 9, http://dx.doi.org/10.1167/tvst.9.2.9. [13]

Lehne, M. et al. (2019), "Why digital medicine depends on interoperability", *npj Digital Medicine*, Vol. 2/1, pp. 1-5, http://dx.doi.org/10.1038/s41746-019-0158-1. [22]

Liang, H. et al. (2019), *Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence*, Nature Publishing Group, http://dx.doi.org/10.1038/s41591-018-0335-9. [44]

Mahajan, A. et al. (2019), "Artificial intelligence in healthcare in developing nations: The beginning of a transformative journey", *Cancer Research, Statistics, and Treatment*, Vol. 2/2, p. 182, http://dx.doi.org/10.4103/crst.crst_50_19. [45]

Masanet, E. et al. (2020), "Recalibrating global data center energy-use estimates", *Science*, Vol. 367/6481, pp. 984-986, http://dx.doi.org/10.1126/science.aba3758. [26]

Matheny, M. et al. (eds.) (2019), *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*, National Academy of Medicine, Washington DC. [14]

Mccall, B. (2020), "News COVID-19 and artificial intelligence : protecting health-care workers and curbing the spread", *The Lancet*, Vol. 2019/20, pp. 2019-2020, http://dx.doi.org/10.1016/S2589-7500(20)30054-6. [10]

Neri, E. et al. (2019), "What the radiologist should know about artificial intelligence – an ESR white paper", *Insights into Imaging*, Vol. 10/1, p. 44, http://dx.doi.org/10.1186/s13244-019-0738-2. [3]

Obermeyer, Z. et al. (2019), "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, Vol. 366/6464, pp. 447-453, http://dx.doi.org/10.1126/science.aax2342. [17]

Oderkirk, J. (2017), *Readiness of Electronic Health Record Systems to Contribute to National Health Information and Research*, http://dx.doi.org/10.1787/9e296bf3-en. [23]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eedfee77-en. [2]

OECD (2019), *Health in the 21st Century: Putting Data to Work for Stronger Health Systems*, OECD Health Policy Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/e3b23f8e-en. [30]

OECD (2019), *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449*. [31]

OECD (2018), *Private Equity Investment in Artificial Intelligence: OECD Going Digital Policy Note*, OECD, Paris, http://www.oecd.org/going-digital/ai/private-equity-investment-in-artificial-intelligence.pdf. [34]

OECD (2017), *Tackling Wasteful Spending on Health*, OECD Publishing, Paris, http://www.oecd-ilibrary.org/docserver/download/8116241e.pdf?expires=1518450288&id=id&accname=ocid84004878&checksum=8647E938E2C1B896ECB03B16256A576B (accessed on 12 February 2018). [1]

OECD (2016), *Recommendation of the Council on Health Data Governance, OECD/LEGAL/0433*, http://legalinstruments.oecd.org (accessed on 7 May 2019). [29]

Pichler, P. et al. (2019), "International comparison of health care carbon footprints", *Environmental Research Letters*, http://dx.doi.org/10.1088/1748-9326/ab19e1. [28]

Rahimian, F. et al. (2018), "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records", *PLoS Medicine*, Vol. 15/11, http://dx.doi.org/10.1371/journal.pmed.1002695. [12]

Reardon, S. (2019), *Rise of Robot Radiologists*, Nature Research, http://dx.doi.org/10.1038/d41586-019-03847-z. [16]

Scavuzzo, J. et al. (2017), *Modeling the temporal pattern of Dengue, Chicungunya and Zika vector using satellite data and neural networks*, Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.23919/RPIC.2017.8211646. [46]

Shadmi, E. et al. (2015), "Predicting 30-day readmissions with preadmission electronic health record data", *Medical Care*, Vol. 53/3, pp. 283-289, http://dx.doi.org/10.1097/MLR.0000000000000315. [5]

Stokes, J. et al. (2020), "A Deep Learning Approach to Antibiotic Discovery", *Cell*, Vol. 180/4, pp. 688-702.e13, http://dx.doi.org/10.1016/j.cell.2020.01.021. [8]

Strickland, E. (2019), *How IBM Watson Overpromised and Underdelivered on AI Health Care - IEEE Spectrum*, IEEE Spectrum, https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care (accessed on 4 March 2020). [18]

Strubell, E., A. Ganesh and A. McCallum (2019), "Energy and Policy Considerations for Deep Learning in NLP", pp. 3645-3650, http://arxiv.org/abs/1906.02243 (accessed on 4 March 2020). [25]

Svoboda, E. (2019), *Your robot surgeon will see you now*, NLM (Medline), http://dx.doi.org/10.1038/d41586-019-02874-0. [4]

Tiwari, P. et al. (2020), "Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation", *JAMA network open*, Vol. 3/1, p. e1919396, http://dx.doi.org/10.1001/jamanetworkopen.2019.19396. [39]

West, S., M. Whittaker and K. Crawford (eds.) (2019), *Discriminating Systems: Gender, Race and Power in AI*, https://ainowinstitute.org/discriminatingsystems.html. [19]

Willyard, C. (2019), *Can AI Fix Medical Records?*, Nature Research, http://dx.doi.org/10.1038/d41586-019-03848-y. [41]

Wolff, J. et al. (2019), "A Systematic Review of Economic Impact Studies of Artificial Intelligence in Healthcare (Preprint)", *Journal of Medical Internet Research*, Vol. 22/2, p. e16866, http://dx.doi.org/10.2196/16866. [37]

Zhavoronkov, A. et al. (2019), "Deep learning enables rapid identification of potent DDR1 kinase inhibitors", *Nature Biotechnology*, Vol. 37/9, pp. 1038-1040, http://dx.doi.org/10.1038/s41587-019-0224-x. [9]

www.oecd.ai
www.oecd.org/health

@OECDinnovation
@OECD_Social

STI.contact@oecd.org
health.contact@oecd.org

OECD
BETTER POLICIES FOR BETTER LIVES