Background paper:
Ethics of AI in global health research

Cape Town, November 2022

Forum abstract
Artificial intelligence (AI) is increasingly being used in global health research but frameworks, policy and best practice for the ethical review and oversight of AI-enabled studies is currently lacking. The Forum will discuss how traditional research ethics regulatory frameworks have responded to the rapid advances in AI technology, and what changes are required, including to the role and responsibility of research ethics committees. It will explore the ethical challenges such as bias, privacy, data provenance and ownership, along with the need for transparency, accountability and engagement during the design and use of AI in global health research. To date, these discussions have predominantly taken place in high-income countries, and low- and middle-income country (LMIC) perspectives have been underrepresented. The Forum will consider the LMIC context where AI has the potential to address critical skills shortages and improve access to care, but where the ethical challenges are made harder due to existing disparities in infrastructure, knowledge and capacity. The Forum will take a multidisciplinary approach to explore how AI technology is being designed and used in health research, reflecting the range of actors involved in this space and the importance of computer scientists and technologists who apply AI for health to understand research ethics frameworks and considerations.

Purpose of this document
This document outlines the scope of the 2022 Global Forum on Bioethics in Research (GFBR) meeting theme and covers the following areas:

1. **Introduction**
2. **Fairness and equity**
3. **Trust and trustworthiness**
4. **Transparency and engagement**
5. **Ethics and regulatory oversight**
6. **Additional governance considerations**

Definitions and scope
The meeting will consider the ethical issues regarding the use of AI systems in global health research. This paper provides a review of the issues related to the topic and points readers to the relevant literature. It is intended as a resource and to provide a foundation of knowledge for colleagues who attend the meeting. The paper also articulates the scope of the meeting and as such is a guide for colleagues who are interested in submitting a case study or governance paper on the meeting theme. The paper addresses issues of theory, substance and process.

Although AI does not have a standard definition, the Organisation for Economic Co-operation and Development (OECD) defines an AI system as:

> "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"[1].

'AI systems' means any AI-based component, software and/or hardware, understanding that AI systems are usually embedded as components of larger systems, rather than being stand-alone systems[2]. An AI system can process far greater quantities of data with which to assess patterns and correlations on a broader scale than would otherwise be possible. Insights from AI can (in theory) support more accurate decision-making by humans, and some systems can act on the insights without human intervention. AI systems can be powered by a number of different techniques, for example, machine learning, deep learning and artificial neural networks. Systems can be either autonomous or semi-autonomous. For the purpose of this paper we refer to 'AI systems' without necessarily specifying the underlying technique.

The AI lifecycle includes a number of steps and a range of actors, including:

- Data collectors and processors
- AI model developers
- AI system deployers
- End-users and stakeholders.

Scientific research has traditionally taken place in academia, driven by the quest for knowledge that is replicable and reliable. However, to date, most research on AI for health care has been driven by private industry with commercial objectives and who's research is not necessarily within the purview of human subject research requirements. Public-private partnerships are also common in this field, for example, where academic research is monetised in the process of transforming research results into a practical AI tool. Whereas strict governance regimes are in place for the multistage testing and validation of commercially developed drugs, these requirements are largely lacking for the commercial development of AI technology for health. Given the prevalence of commercial interests in this space, the Forum scope includes global health research by private industry and public-private partnerships.

Possible uses of AI systems in health research include[3]:

- **Basic research** e.g. drug discovery, protein folding predictions, use in genomics, vaccine development
- **Clinical research** e.g. to develop AI-based tools for screening and triage, diagnosis, prognosis, decision-support and treatment recommendation, and to manage clinical trial design and conduct, for example pre-screening and identifying suitable patients and analysing trial data in real-time
- **Public health research** e.g. AI-based tools to monitor and predict the spread of an epidemic or monitoring and assessing population health, and targeting public health interventions
- **Health systems research** e.g. to assess and refine delivery and access to health services.

For each of these health research purposes, AI could have the role of an:

[1] OECD/LEGAL/0449 (2019) Recommendation of the Council on Artificial Intelligence. Though some have cautioned against describing AI systems as 'autonomous'. See: Priest, C. (2021) Humans and AI: Should we describe AI as autonomous? Data Robot Blog and Totschnig, W. (2020) Fully Autonomous AI. *Sci Eng Ethics* 26:2473–2485

[2] For a comprehensive account of the AI techniques and sub-disciplines that are currently used to build AI systems see: The European Commission's High-Level Expert Group on Artificial Intelligence (2018) A definition of AI: main capabilities and scientific disciplines

[3] Use of AI across these different fields of research varies. For example, it is currently more commonly applied in basic research, in comparison to clinical research.

- "**algorithm for discovery**" (e.g. being used as a method to generate hypotheses or answer research questions such as discovering associations in population health data that reveal a new disease group or discovering potential drug candidates) or
- "**algorithm for intervention**" (e.g. being used as a component of an intervention, the impact of which is being examined in a particular setting such as a clinical screening tool which generates an output for human consideration and action).

Despite these different uses and roles, AI systems are grouped for the purpose of this paper because they share a number of ethical issues regarding data provenance and ownership, privacy and security, the potential for biases, the need for transparency, accountability and engagement and questions about whether current governance structures are fit for purpose. The nature and scale of these ethical and governance issues will depend on the specific health research use and context. The upcoming GFBR meeting will provide an opportunity for stakeholders (e.g. bioethicists, researchers, computer scientists, funders, policy-makers, advisory board members) to engage in rigorous critical assessment of these ethical issues through discussion of real-life LMIC global health research case studies.

This paper is being published with the call for applications for participants to attend the Forum. GFBR is seeking three types of participants for the meeting[4]:

1. **Case study presenters** will present their research experiences and the ethical issues that have emerged regarding the use of AI in health research in LMIC settings
2. **Governance paper presenters** will present on topics such as regulation, policies, guidance, tools and issues associated with ethics and other review mechanisms
3. **Participants** will attend the meeting and actively take part in plenary and small group discussions and networking opportunities.

Case studies

The GFBR organisers are looking for interesting and important cases that are relevant to the theme. The cases could demonstrate the **development of good practice; highlight ethical challenges; demonstrate situations in which ethical practice failed; or present unresolved questions** for the global community. The organisers encourage cases that address research ethics oversight and cases that address past GFBR topics[5]. However, the actual topics considered at the meeting will be defined by the case studies that are submitted. In this way, GFBR aims to be responsive to applicants and the issues that they consider most important.

Case studies should focus on the ethical issues that **result from the use of an AI system in global health research**. Case studies on research that happens to use an AI system but the ethical issues which relate to an aspect of the study that are not explicitly tied to the use of AI are not in scope. For example, a case about the large scale and variety of health data required to train an AI system, which may increase privacy and security concerns, would be a stronger case than one that discusses data sharing issues that are more common to other types of non-AI research.

The scope includes case studies that e.g.:

- Focus on **issues around conducting health research in LMICs**[6].
- Are from **any stakeholder perspective**, including ethicists, policy-makers, researchers, clinicians, computer scientists, and healthcare workers.

[4] See the call for applications for details on how to apply and the requirements for each type of participant.
[5] For example, mental health research, genomics, research during epidemics, novel trial designs (e.g. adaptive trials). For the full list of past topics see: www.gfbr.global/past-meetings.
[6] However, we do not want to exclude case studies from high-income countries if there could be valuable lessons to learn, and some parallel or relevant ethical considerations. If your case study relates to a high-income country please use the commentary section to draw-out the relevance for research in LMICs.

- Are from **any organisational perspective** e.g. academic, technology companies, government, non-governmental organisation (NGOs) and public-private partnerships.
- Address the **ethical issues associated with the lifecycle of developing, validating and using an AI system in the health research context**. This could include:
    - Model development:
        - Collecting and processing data on which to train the AI algorithm
        - Designing and developing the algorithm
        - Training the algorithm (e.g. using a 'test set' and 'tuning set' of data)
    - Model validation:
        - Using an internal or external test set to validate the algorithm
    - Model use in health research. For example:
        - Assessing the impact of an "**algorithm for intervention**" e.g. a prospective observational trial or an interventional clinical trial to evaluate an AI-based clinical tool in a clinical setting, which could include an assessment of how use of the algorithm may change the outcome for the patient, the behaviour of physicians or the patient/ physician relationship.
        - Using a validated "**algorithm for discovery**" in research e.g. to generate hypotheses or answer research questions using population health data.
- In addition, GFBR is open to case studies of research on AI to develop systems with health applications e.g. projects that may be framed as data or computer science developed by technology companies and which are not necessarily characterised as 'health research' or subject to research governance requirements.

Governance papers

The scope includes governance papers that e.g.:

- Focus on **institutional, national, regional or international regulation, guidelines, policy, principles or codes of practice**. The paper should speak specifically to the relevance and impact of these document(s) on AI in health research.
- Present **issues and initiatives associated with research ethics review** (e.g. research ethics frameworks and procedures, components of the review, role and skills of RECs) or **technology review, privacy review**, etc. that may take place in parallel to the REC review process.
- Discuss other **governance bodies, mechanisms or tools** (e.g. advisory councils or committees, impact assessments, data sharing or data use policies, research reporting standards).

The governance paper can be either:

- **Practical** (e.g. discuss gaps in national regulation or issues with research ethics review processes and propose a practical solution such as a new tool or mechanism) or
- **Theoretical** (e.g. draw on good theory about public-private partnerships in AI health research).

Key themes and questions

We indicate below some examples of issues considered important by the organisers. These are not exhaustive and are intended only as examples. Case studies and governance papers should focus on research in LMICs, though examples from high-income countries (HIC) will be considered if they show relevance to LMIC settings. Case study and governance papers could address (but are not limited to) one or more of the following questions. Where the questions refer to a 'researcher' this includes academic, commercial and government sector and includes those who develop AI systems and those who validate and use them in health research.

Fairness and equity

- What processes, tools and checks are available to researchers to mitigate and identify data and algorithm bias? Who should be involved in assessing bias and issues of equity during the development and use of AI systems in health research?

- What challenges are faced by researchers in LMICs in developing and managing equitable international collaborations in the field of AI-based health research (with both public and private organisations)? What solutions have been proposed?

- How can ethics and the social sciences be embedded to inform the technical design and development of AI for health research and to mitigate potential unforeseen risks? Are there examples of best practices?

- What opportunities or initiatives are there for increasing the collective leverage of LMICs on data ownership to enhance greater access to training data for AI and to stimulate locally and globally driven AI health research (e.g. shared data platforms, algorithm registers)?

- How can inclusion be promoted during the development and use of AI for health research and what could this look like (e.g. at the level of including a range of different stakeholders and/ or different cultures and perspectives and through training in the ethics of AI health research in LMICs)?

Trust and trustworthiness

- To what extent do current practices for using AI in health research – which were largely developed in HICs – resonate with the culture and values of stakeholders in LMICs (e.g. with respect to how personhood and privacy are conceived)?

- How can relevant values and perspectives in specific LMIC settings be identified and incorporated to foster and ensure ethical design of AI and the prioritisation of research that is most relevant to those settings?

- Are there unique features of AI health research that demand new approaches to consent, privacy and security (e.g. auditable e-consent or broad consent processes)? What new approaches have been used and what issues can weaken these approaches (e.g. power imbalances between data collectors and those who provide data)?

- How do design issues for e-consent for AI-enabled research impact the role of consent as a safeguard of autonomy? Are there examples of successful designs?

Transparency and engagement

- How does the use of AI in different research settings influence or change the way researchers should think about doing engagement and what practical approaches have been proposed or tested (e.g. basic research vs clinical research vs population research)?

- Who should be engaged during health research that uses an AI system (e.g. representatives from marginalized groups, local patient populations, communities more broadly etc.). When should they be engaged, how and for what purpose (e.g. for setting priority topics to explore, to inform the design of the algorithm and research, to help identify and mitigate unforeseen risks etc.)?

- What tools and criteria are being used to assess the impact of AI algorithms (e.g. on equity, privacy, human rights and safety)? Should assessment and certification take place during the research process, after deployment or some combination of both, and which stakeholders should be involved?

- To what extent does use of the tool (e.g. an algorithm impact assessment) exhaust a researcher's ethical responsibility? If not, what else is required?

- What are the roles and responsibilities of stakeholders (e.g. researchers, funders, policy-makers, private industry, journals etc.) in facilitating ethical and equitable development and transparent reporting of AI-based health research? Are there examples of best practice?

Ethics oversight

- What are current practices in ethical review of AI research and ensuring their ethical conduct within and across countries and settings?

- Specifically, what challenges have RECs faced when reviewing AI-based health research protocols and how were these challenges overcome (e.g. aspects of consent, risk/benefit assessment, privacy concerns, complexity of algorithms etc.)?

- How should traditional research ethics regulatory frameworks be adapted to respond to AI-based health research? (e.g. Should ethics review extend beyond the initial phase and also address other parts of the AI lifecycle? Is new guidance required to re-define the scope of REC review, extending it from the traditional protection of individual interest to also consider and balance societal benefits and risks?)

- How should traditional research ethics procedures be adapted to respond to AI-based health research? (e.g. Should algorithmic impact assessments or other reporting metrics be part of the ethics review process or parallel complementary reviews? In what ways can RECs acquire the necessary expertise to review AI health research – training, expert input etc?)

Governance

- What are the current governance structures and processes to support AI-based health research? Are they sufficient or are other governance mechanisms required?

- Which models do different countries use to govern AI-based health research (e.g. self-regulation, regulation, guidance). Does this depend on the type of health research application?

- How can ethical principles be implemented in practice and what methods, processes, and frameworks can be used or are needed for researchers who use AI to better understand and operationalise ethics within their own research?

- What is required for market authorisation of a new AI-based application and to what extent does this align (or not) with research ethics review requirements?

1. Introduction

Interest in AI has grown enormously in recent years given its potential as an enabling tool with applications in all aspects of life (e.g. healthcare, banking, agriculture, social media). In the field of health, new AI technology has the potential to improve diagnosis, treatment, drug development,

health systems and public health functions. Several global initiatives have addressed the opportunities of AI in the health sphere, and recognised the development, use and governance challenges:

- The **World Health Organization** (WHO) has highlighted the importance of digital technologies to help increase universal access to affordable person- and community-centred care and services[7]. Its 2021 guidance on *Ethics & Governance of Artificial Intelligence for Health* identifies the ethical challenges and risks with the use of AI for health. The guidance provides six consensus principles to ensure AI works to the public benefit of all countries and contains a set of recommendations for the governance of AI[8].
- The **World Economic Forum** initiative on AI is broad in scope[9], and includes health projects e.g. supporting the development of new data laws and a national AI policy in Rwanda to underpin the use of AI tools to improve healthcare[10] and publishing a framework on the use of ChatBots in healthcare[11].
- The **OECD** works to support governments by measuring and analysing the economic and social impacts of AI technologies and applications, and engaging with stakeholders to identify good practices for public policy[12]. It has published papers on 'Laying the foundations for AI in health'[13] and 'Trustworthy AI in health'[14].

These initiatives, and others, specifically address – or at least are relevant to – health research. However, the picture is not straightforward given the diffuse nature of AI and because the speed of adoption has outpaced AI's governance producing several 'grey areas':

- **What counts as health research?** For example:
    - An AI system used by an academic to analyse the spread of a pandemic may be considered research but use of a similar AI system by government for the same purpose may be considered public health monitoring
    - The distinction between 'evaluating' or 'piloting' a new AI system, and conducting research into how the AI system works
    - Using user-derived data from social media to train an AI algorithm and develop a health-related app.

- **In turn, what then is required in terms of governance?** For example:
    - Some AI development may be characterised as computer science which traditionally has not been subject to research governance requirements or considered human subject research thereby requiring research ethics review
    - Regulation in some countries may apply to publicly funded health research only and not to AI systems developed by technology companies. Product research carried out by technology companies has, historically, not had to follow the same rules to protect research participants (e.g. requirements for REC review).

While there has been an escalation of AI research in recent years it has been compromised by ethical challenges (e.g. relating to privacy, bias, accountability and governance). The field has also been complicated by the number and types of actors involved in the AI lifecycle and the different cultures they bring. Where once health research was performed primarily by academic researchers,

[7] World Health Organization (2017) Meeting report: Big data and artificial intelligence for achieving universal health coverage: an international consultation on ethics
[8] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance
[9] World Economic Forum Artificial Intelligence website
[10] World Economic Forum (2022) How Rwanda's vision for data and AI is revolutionizing its services and healthcare system
[11] World Economic Forum (2021) Chatbots RESET framework pilot projects: Using chatbots in healthcare
[12] OECD Artificial intelligence website
[13] Hashiguchi, T.C.O., Slawomirski, L. & Oderkirk, J. (2021) Laying the foundations for artificial intelligence in health. *OECD Health Working Papers*, No. 128, OECD Publishing, Paris
[14] OECD (2020) Trustworthy AI in health, Background paper for the G20 AI Dialogue, Digital Economy Task Force

the field of AI for health includes computer scientists, technology companies, start-ups and non-profit organisations. The nature and scale of health data used in research has also changed with the advent of granular personal health-related data generated in people's daily lives.

Research ethics for AI systems intended for clinical use

McCradden *et al*.[15] discuss the current gap between the development of a robust algorithm and its clinically meaningful application, asserting this is "epistemic (methodological) and ethical, generated by a clash between the cultures of computer science and clinical science". They note that few prospective, controlled studies of clinical machine learning (ML)[16] models have been conducted and call for the responsible evaluation, validation, and clinical integration of clinical ML using sound research methods, as seen in other areas of clinical research. They propose a comprehensive research ethics framework that can apply to ML research across its development cycle, consisting of three stages:

- **Exploratory ML research**: hypothesis generation and model development utilising and comparing multiple computational techniques to explore retrospective data for models with potential clinical applicability
- **Silent period evaluation**: hypothesis testing and clinical validation using a prospective observational trial where outputs are not visible to the clinical team
- **Prospective clinical evaluation** using an observational, quasi-interventional, or interventional clinical trial.

According to McCradden *et al*., "this pathway can accommodate many research designs from observational to controlled trials, and the stages can apply individually to a variety of ML applications". The paper provides a detailed explanation of the framework and is essential reading for those interested in this year's GFBR theme.

AI in global health research

AI systems have significant potential and importance in LMICs for health as a way of[17, 18, 19]:

- addressing critical medical skills and staff shortages e.g. by supporting task-shifting through the empowerment of nurses and community healthcare workers
- delivering services previously requiring scarce medical officers and improving access to services in remote areas e.g. using AI-powered tools such as mobile phone apps to reach people directly with targeted health campaigns
- getting more value out of available data and addressing gaps in representativeness of existing health data and research.

AI technology is more easily transferable to LMICs than, for example, technologies required for vaccine development or pharmaceutical development, meaning adoption in LMICs has the potential

[15] McCradden, M.D., Anderson J.A., Stephenson E.A. *et al*. (2022) A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics* 22(5):8-22
[16] We refer specifically to ML models to reflect the terminology used by McCradden et al.
[17] Williams, D., Hornung, H., Nadimpalli, A. *et al*. (2021) Deep learning and its application for healthcare delivery in low and middle income countries. *Front. Artif. Intell.* 4
[18] Parry, C.M. & Aneja, U. (2020) Chatham House research paper: Artificial intelligence for healthcare: Insights from India
[19] Verma, A., Rao, K., Eluri, V., *et al*. (2020) Building a collaborative ecosystem for AI in healthcare in low and middle income economies. Atlantic Council website

to be quite fast[20]. Uptake is already being accelerated by technology companies in HICs setting-up, or partnering, with organisations in LMICs[21].

However, health-focused AI systems have mostly been developed in HICs and as such there are few robust and contextualized evaluations that can guide informed decision-making in LMIC contexts[22]. This situation gives rise to several risks and challenges, including the risk of biased or defective results if the data used to train the algorithm is not representative or generalisable to the population on which it will be applied in the LMIC context[23]. In addition, technology will not be fit for purpose and could cause more problems, if local institutions and people are not upskilled with the AI knowledge[24] required for them to understand the AI system and its applicability and potential impact in their context. Further challenges for implementation of AI systems in LMICs – and HICs – remain, including potential algorithmic bias due to unrepresentative data even when local data is used, oversight, and capacity to audit systems locally.

The applicability of HIC-developed AI systems to LMICs may also be limited due to differences in culture that challenge the assumptions and the premise on which AI is built. For example, varying notions of personhood in different global contexts may influence opinions and approaches to privacy and data-sharing in AI health research. Indeed, the very nature of AI, built on a traditional Western view of personhood based on rationality, has been criticised as limited due to its lack of proper context and relationality, which put it at risk of perpetuating racial and gender biases[25].

To date, the landscape of AI ethics guidelines has largely been occupied by Western countries which, it has been argued, denies the involvement of a fuller range of cultures, the variety of normative perspectives and, ultimately, the true complexity of ethical analysis[26]. Given that AI inherently interacts with its surroundings and the cultural, political and environmental context there is a need to understand how AI systems used in health research may impact or be accepted by society in various regions around the world[27]. And a contextualised approach to the ethics and governance challenges is needed to take account of different world views and philosophies[28].

2. Fairness and equity

A number of practical challenges to implementing AI systems in LMICs may impact on their fair and equitable uptake and distribution (Box 1). The AI ethics literature often cites 'fairness' as a key principle. While there are variations in interpretation, it often relates to the need to avoid algorithmic bias in input data, modelling and algorithm design[29].

[20] Subject to the challenges listed in Box 1 below relating to power, internet, mobile technology infrastructure etc. all of which may perpetuate existing inequalities in access to such technologies and thus potentially to health outcomes
[21] Dean, J. & Cisse M. (2018) Google AI in Ghana. Google Blog
[22] Alami, H., Rivard, L., Lehoux, P. *et al.* (2020) Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Global Health* 16
[23] Carrillo-Larco, R.M., Tudor Car, L., Pearson-Stuttard, J., *et al.* (2020) Machine learning health-related applications in low-income and middle-income countries: a scoping review protocol. *BMJ open* 10(5)
[24] OECD defined AI knowledge as '… the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle'
[25] Mhlambi, S. (2020) Carr Center Discussion Paper: From rationality to relationality: Ubuntu as an ethical & human rights framework for artificial intelligence governance.
[26] Goffi, E.R. (2021) The importance of cultural diversity in AI ethics. Institut Sapiens website
[27] Responsible AI network Africa. Institute for Ethics in Artificial Intelligence website
[28] Mhlambi, S. (2020) Carr Center Discussion Paper: From rationality to relationality: Ubuntu as an ethical & human rights framework for artificial intelligence governance
[29] González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2)

Box 1: Challenges to implementing AI in LMICs

Data, infrastructure and hardware:

- Availability, accessibility and quality of data
- Access to reliable and affordable internet
- Lack of access to sufficient computing power
- Digital inclusion and connectivity: device access, ownership and capability
- Unreliable power infrastructure

Human capital, funding and other constraints:

- Human capital, education and skills
- Lack of investment
- Poor transferability
- Automation and the risk of job losses

Extract from: Sharma, A., Ajadi, S. & Beavor, A. (2020) GSMA report: Artificial intelligence and start-ups in low- and middle-income countries: Progress, promises and perils

Bias in training data and algorithm failures

The potential benefits of AI using biomedical big data are ethically important[30]. AI can be used to analyse and identify patterns in large and complex datasets faster and more precisely than has previously been possible[31], but it may also deepen inequalities by exacerbating health disparities.

Bias in data sets may arise due to poor data collection methods and/or due to individual or cultural values or biases (gender, socio-economic status, caste etc.). Where training data is biased this can lead to unrepresentative data sets, impacting on how algorithms are developed, trained and used, thereby further impacting on how the results are interpreted and their value (or not) for certain populations. Data bias and algorithmic harm are potentially compounded by the scalability, power and homogeneity of the AI system which amplifies its effects[32]. Populations in LMICs can be particularly vulnerable to bias and fairness in AI systems, due to a lack of technical capacity, existing social bias against minority groups, and a lack of legal protections[33].

On the other hand, AI has positive potential to mitigate existing bias within healthcare systems e.g. reducing human error and reducing biases that may be present within healthcare research and public health databases[34].

Mechanisms are needed to mitigate and identify inbuilt biases and assess their impact. This could include using training data that is representative of the population in which the AI system will be used, increasing diversity among the people who label data and validate algorithms and through

[30] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance

[31] Nuffield Council on Bioethics in Research (2018) Bioethics briefing note: Artificial intelligence (AI) in healthcare and research

[32] Alami *et al*. explain that a medical error generated by an AI application could affect a large number of people at the same time, whereas, traditionally, an error made by a clinician affects only a smaller number of persons: Alami, H., Rivard, L., Lehoux, P. *et al.* (2020) Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Global Health* 16. Campbell *et al*. explain the issue of homogeneity and that if an AI system performs poorly on a certain disease, a certain population, or both, this effect may be replicated around the world. By contrast, human decision makers may be biased, but the effect may be mitigated, to some extent, by their diversity of biases: Campbell, J.P., Lee, A.Y., Abràmoff, M., *et al.* (2020). Reporting guidelines for artificial intelligence in medical research. *Ophthalmology* 127(12)

[33] Fletcher, R.R., Nakeshimana, A. & Olubeko, O. (2021) Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front. Artif. Intell.* 3

[34] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)

human oversight to ensure AI is not unfairly biased[35]. Campbell *et al*. propose that it is incumbent on researchers to explore algorithm failures within their available data, understanding that input parameters will likely be higher in clinical practice than in a tightly regulated clinical trial[36]. Equity assessment have also been proposed as a core component of regulatory approval that: "(1) require attention to health disparities and the potential for group harms as part of the clinical evaluation process in premarket review and (2) continue this focus on disparities and group harms through a "health equity review" that is part of postmarket, real-world performance monitoring" as an inclusive process that includes multiple stakeholders[37].

Transferability, data scarcity and quality

Algorithms are subject to the quality of the initial data imputed into the AI system[38]. In this sense, inequity can arise due to the poor transferability of AI algorithms to the LMIC context (e.g. where the training data derived from HIC is unrepresentative). Algorithms themselves could also potentially be unrepresentative if they follow decision logic that aligns with a particular epistemological view of the world.

Given that many disease profiles and prevalence are unique to low-resource settings, targeted AI applications based on locally sourced data can potentially improve the performance of the AI for the local population[39]. Sallstrom *et al*. argue that any import of foreign AI technology must be done with awareness of its development process and limitations and local datasets should be made available to local researchers and companies working with imported AI technology, to ensure locally applicable outcomes[40].

However, there may be data scarcity in LMICs due to challenges around collecting data from individuals without financial or geographic access to health services[41]. Other challenges articulated by the Nuffield Council on Bioethics that apply to both LMIC and HIC include inconsistencies in the availability and quality of data, medical records not being (consistently) digitised across health systems, and the lack of interoperability and standardisation in health systems, digital record keeping, and data labelling[42].

'Data commons' and 'open digital commons' are being explored as a framework to develop a public AI utility where global stakeholders can enhance and utilize data sets, libraries of software and common tools[43, 44]. Verma *et al*. provide a comprehensive account of open source tools and their potential for enabling AI healthcare innovations in LMICs including data de-identification tools, data-banks, annotation tools, collaborative spaces and platforms and peer review platforms[45]. It is noted

[35] Nuffield Council on Bioethics in Research (2018) Bioethics briefing note: Artificial intelligence (AI) in healthcare and research
[36] Campbell, J.P., Lee, A.Y., Abràmoff, M., *et al*. (2020) Reporting guidelines for artificial intelligence in medical research. *Ophthalmology* 127(12)
[37] Ferryman, K. (2020) Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. *Journal of the American Medical Informatics Association* 27(12)
[38] Brummel, E.S. (2017) Confronting natural conflicts of interest and artificial intelligence. *Journal of Law and the Biosciences* 4(2)
[39] Williams, D., Hornung, H., Nadimpalli, A. *et al*. (2021) Deep learning and its application for healthcare delivery in low and middle income countries. *Front. Artif. Intell.* 4
[40] Sallstrom, L., Morris, O. & Mehta, H. (2019) ORF issue brief: Artificial intelligence in Africa's healthcare: ethical considerations
[41] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)
[42] Nuffield Council on Bioethics in Research (2018) Bioethics briefing note: Artificial intelligence (AI) in healthcare and research
[43] Project resilience, International Telecommunication Union website
[44] Goldstein, E., Gasser, U. & Budish, R. (2018) Data Commons Version 1.0: A framework to build toward AI for Good. *Medium*
[45] Verma, A., Rao, K., Eluri, V., *et al*. (2020) Building a collaborative ecosystem for AI in healthcare in low and middle income economies. Atlantic Council website

that several countries have detailed such data-banks, including India's proposal for a National AI Platform[46].

Language has been recognised as a barrier to health research[47] and is also pertinent to the transferability of AI systems. English is fast becoming the standard language for AI applications and translations in other languages can pose problems. Whereas in other types of research, trained researchers can normally mediate technologies, this is not always the case with AI driven research developed in an unfamiliar language (e.g. given the added complexity of the underlying data that trained the AI being labelled in the unfamiliar language).

Data ownership and management

The collection, use, storage, and sharing of both individual and population-based health data raise important questions in terms of consent, ownership, and access[48, 49]. AI systems are fundamentally built on data and the ethical issues around their use reflect debates about data sharing in health research. Data sharing has been recognised for its potential to increase scientific efficiency by maximising the availability and utility of data and is something that research funders and journals are increasingly promoting to improve the transparency and utility of research, with the ultimate aim of improving health[50]. However, data sharing has the potential to exacerbate existing inequalities, particularly if data sharing benefits only those from well-resourced institutions, leaving researchers in low-resourced settings worse off. In the context of AI and data provenance, the divide between those who accumulate, acquire, analyse and control data and those who provide the data but have little control over their use has been recognised[51].

Power dissymmetry in international research partnerships can result in inequities in the research process. For example, in relation to:

- **Opportunities for local researchers to influence the research agenda** so research responds to issues of local importance as well as global concerns
- **Recognition for primary data collectors** (including in the publishing process) and recognising all intellectual contributions to the research process (e.g. data analysts)
- **Data being available via appropriate access mechanisms** to the countries and organisations that provided the data, and to others more broadly
- **Whether or not there is local investment** in digital infrastructure, research capacity, training and infrastructure to ensure that the products of AI and big data are also generated by researchers and companies in LMIC[52].

Although earlier debates about data sharing are relevant, the nature of AI presents fundamental new challenges due its scale, potential impact and the range of actors involved. Whereas health research would traditionally use data from health records, new sources of personal health data are being generated and acquired from social media and direct-to-consumer wellness products and

[46] Ministry of Electronics and Information Technology, Government of India (2021) Report of Committee: On platforms and data on artificial intelligence
[47] Ransing R., Vadivel R., Halabi S.E., *et al.* (2021) Language as multi-level barrier in health research and the way forward. *Indian Journal of Psychological Medicine*
[48] Alami, H., Rivard, L., Lehoux, P. *et al.* (2020) Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Global Health* 16
[49] It is beyond the scope of this paper to describe regulatory requirements for research use of anonymised/non-personal data vs pseudoanonymised/de-identified data but the importance of this distinction, and the spectrum of identifiability, is acknowledge. For a discussion on this topic see: Ferretti, A., Ienca, M., Sheehan, M. *et al.* (2021) Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics* 22(51)
[50] Global Forum on Bioethics in Research (2019) Meeting report: Ethics of data sharing and biobanking in health research
[51] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance
[52] *ibid*

mobile apps[53]. The vast amounts of digital data are changing health research and healthcare as data-driven AI technologies are developed for integration into the health sector via untraditional research processes and sometimes without traditional research safeguards. For example, when consent is not obtained for the training and evaluation phase of an AI technology (e.g. a mobile health app). These issues are picked up later in this paper.

The presence of technology companies and public-private partnerships in the field of AI for health is significant. Google's AI Health & Bioscience research area covers genomics, public health and imaging and diagnostics[54]. Likewise, Meta is developing AI tools to support MRI imaging[55], COVID-19 forecasting[56] and to help researchers better represent and analyse cellular development[57]. Large technology companies are supporting AI infrastructure in LMICs[58], and funding other organisations to conduct projects on AI for health[59] and AI ethics[60] in different regions of the world.

Some public-private partnerships have been controversial, in terms of how data were acquired and used and the justification for their use. In a well-reported case, identifiable data from 1.6 million patients were shared by the Royal Free London NHS Foundation Trust (UK) with Google DeepMind, an AI research company, with the intention of improving the management of acute kidney injuries with a clinical alert app. Murphy *et al*. highlight the questions of whether the quantity and content of the data shared was proportionate for the intended use, and why it was necessary for DeepMind to retain the data indefinitely[61]. They also note the "absence of adequate patient consent, consultations with relevant regulatory bodies, or research approval, threatening patient privacy, and consequently public trust". A detailed account of the partnership by Powles and Hodson conclude that "from the perspective of patient autonomy, public value, and long-term competitive innovation, existing institutional and regulatory responses are insufficiently robust and agile to properly respond to the challenges presented by data politics and the rise of algorithmic tools in healthcare"[62].

Data colonialism

Current discourse on data colonialism encompasses the geopolitical context – and contest – to be a global leader in AI and the global commodification of human experience where the human is substituted as an assemblage of their data points[63]. The notion of "data mining" and its colonial connotations has been described as symbolic of the extent to which the human experience (turned into data) is perceived by some as a raw material free for the taking[64], sometimes without due regard for informed consent, privacy and autonomy. WHO has pointed out that even where informed consent may be acquired, it may be insufficient to compensate for the power dissymmetry between the collectors of data and the individuals who are the sources[65].

[53] Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17, 137
[54] Google Research Health Team website
[55] MetaAI fastMRI website
[56] MetaAI (2021) Blog: Using AI to help health experts address the COVID-19 pandemic
[57] MetaAI (2020) Blog: Poincaré maps: Hyperbolic embeddings to understand how cells develop
[58] Dean, J. & Cisse M. (2018) Blog: Google AI in Ghana
[59] These projects address AI for social good broadly with some focussing specifically on health: Google (2018) Google AI impact challenge. Working together to apply AI for social good
[60] These projects address AI ethics broadly with some focussing specifically on health: Andrade, N. (2020) Meta Research website: Promoting AI ethics research in Latin America and the Caribbean; Meta Research (2019) Ethics in AI research awards - India; Meta Research (2020) Blog: Facebook announces award recipients of the ethics in AI research initiative for the Asia and Pacific
[61] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)
[62] Powles, J. & Hodson, H. (2017) Google DeepMind and healthcare in an age of algorithms. *Health Technol.* 7:351–367
[63] Adams, R. (2021) Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews* 46(1-2):176-197
[64] Birhane, A. (2019) Blog: The algorithmic colonization of Africa
[65] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance

Mollura *et al*. recognise the health equity need to share data as without it, AI may not be created for LMIC populations. However, challenges raised by international data sharing are exacerbated by the lack of regulatory frameworks for data rights. They describe their approach to addressing this problem in global health radiology which includes the creation of the RAD-AID Friendship Data Trust. This platform offers resource-poor health institutions the opportunity to contribute anonymized data to a not-for-profit collective data trust to engage AI developers willing to make AI software available pro bono. As they describe, this increases collective leverage of LMICs on data ownership and procures greater access to AI while also stimulating locally and globally driven AI development[66].

The AI colonialism debate also relates to the ethical rules of AI, which are largely developed in HICs, to the exclusion of broader ethical and socio-cultural perspectives[67, 68].

3. Trust and trustworthiness

There is a need to build trustworthy data acquisition and AI development processes, paying due regard to consent, privacy, security, safety, reliability and utility for communities. These aspects are addressed in the following section, along with issues of reproducibility and conflict of interest.

Consent, privacy and security

The ethical and governance challenges associated with sharing and using big data in health research are well documented[69], and can be heightened in the context of AI given the large amounts and varieties of data required to train AI algorithms. Issues of consent, privacy and trust cut across the full range of data sharing and use scenarios but different starting points bring different ethical questions[70]. For example, the questions will depend on the data type, how data was collected, from whom, by whom, for what purpose(s) and jurisdictional differences regarding what data falls under data privacy and research regulatory requirements.

In the context of autonomous AI systems being trained on patient data, Abramoff *et al*. argue that data used by the system developers should be traceable to an authorisation and that transparency on the part of the system developers, through written agreements, is essential to assess whether individuals have adequately authorised data use[71]. This raises a question about the understandability of such agreements and the implications for informed consent. A related issue is how much information should researchers disclose to research participants about how the technology works, and how a researcher can fulfil their disclosure obligation regarding this aspect in a manner that most participants will likely understand.

Abramoff *et al.* also recommend that authorisation processes are auditable with "security controls to ensure that data are being used in accordance with the scope for which such use was authorised and to protect the data from unauthorised use or access"[72]. However, determining which data uses are permitted for a given purpose is not always easy[73] and – as WHO points out – true informed

[66] Mollura D.J., Culp M.P., Pollack, E., *et al*. (2020) Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology* 297(3):513-520
[67] Goffi, E.R. (2021) The importance of cultural diversity in AI ethics. Institut Sapiens website
[68] Mollura D.J., Culp M.P., Pollack, E., *et al*. (2020) Artificial intelligence in low- and middle-income countries: innovating global health radiology. Radiology 297(3):513-520
[69] Vayena E. & Blasimme A. (2017) Biomedical big data: new models of control over access, use and governance. *J Bioeth Inq.* 14:501–513
[70] Global Forum on Bioethics in Research (2019) Meeting report: Ethics of data sharing and biobanking in health research
[71] Abràmoff M.D., Tobey D. & Char D.S. (2020) Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process. *Am J Ophthalmol* 214:134-142
[72] *ibid*
[73] Vayena E., Blasimme A. & Cohen I.G. (2018) Machine learning in medicine: Addressing ethical challenges. *PLoS Med* 15(11)

consent is increasingly infeasible in an era of big data, especially in an environment driven mainly by companies seeking to generate profits from the use of data[74].

Murdoch highlights a number of consent and privacy issues that arise in the commercialisation of AI systems that use patient data[75]:

- Technologies are often made in an academic research environment, but then undergo a commercialisation process involving private entities. This could have implications for the research consent process and what participants are told if private companies ultimately obtain the patient data, and become responsible for utilising and protecting it.
- Sometimes, patient data has been moved from the jurisdiction from which it was obtained and onto servers owned by private companies.
- Even de-identified data could be re-identified by external parties using complex algorithms. As data are collated and stored during the training and use of algorithms, the risk and potential severity of security breaches grows[76]. This is a particular concern for health data that tends to contain more sensitive higher risk personal information. This security risk also applies to non-commercial uses of AI systems.

The need for new and improved forms of data protection and anonymization has been highlighted, along with the potential to use generative models that develop the ability to generate realistic but synthetic patient data with no connection to real individuals[77]. 'Pooling insights' rather than pooling patient data has also been proposed as a novel solution in which algorithms are designed to reinforce each other in their collective analyses without exchanging data[78].

Privacy is broadly conceived of as an individualised right in many Western countries but may not always be the main data-related value in more communitarian-centred societies[79]. Reviglio and Alunge highlight that "transplanted ethical norms and values can collide with those of the communities in which [AI systems] are deployed". They suggest that Ubuntu – like other communitarian moral philosophies – can strengthen the development of a more relational conceptualization of privacy which could enrich and develop the current paradigm of privacy protection to address the challenges of AI[80]. Mhlambi has developed, and described in details, the use of Ubuntu as an Ethical and Human Rights Framework for AI governance[81].

Some AI-based health research utilises user-generated data derived from apps, digital devices and social media. Research has suggested that such social data can act as a predictor for depression, suicide risk factors, mood changes and flu outbreaks but often the people from whom the data derived are not informed – or asked to consent – to the research use[82].

Pickering uses a COVID-19 contact-tracing app to illustrate the potential confusion that can arise when consent is requested in this context: A user may sign-up to the app, entering into a contract with the service provider via their Terms of Use. Consent may be requested by the provider as the

---

[74] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance
[75] Murdoch, B. (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 22, 122
[76] Shaw, J., Rudzicz, F., Jamieson, T. & Goldfarb, A. (2019) Artificial intelligence and the implementation challenge. *Journal of medical Internet research 21*(7)
[77] Murdoch, B. (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 22, 122
[78] Peumans, P., Verachtert, W. & Wuyts, R. (2021) How AI advances can enable medical research without sharing personal data. World Economic Forum website
[79] Gillwald, A. & Adams, R. Artificial intelligence carries a huge upside. But potential harms need to be managed. (2021) *The Conversation*
[80] Reviglio, U. & Alunge, R. (2020) "I am datafied because we are datafied": an Ubuntu perspective on (relational) privacy. *Philos. Technol.* 33:595–612. Also referenced in: Aggarwal, N. (2020) Introduction to the special issue on intercultural digital ethics. *Philos. Technol.* 33:547–550
[81] Mhlambi, S. (2020) Carr Center Discussion Paper: From rationality to relationality: Ubuntu as an ethical & human rights framework for artificial intelligence governance
[82] Norval C. & Henderson T. (2020) Automating dynamic consent decisions for the processing of social media data in health research. *J Empir Res Hum Res Ethics* 15(3):187-201

legal basis for processing the personal data. In turn, the data may be shared with researchers, requiring a different consent appropriate to the research purpose. They argue that "conflating data protection consent and research ethics consent may confuse the data subject/participant as well as restrict what a researcher can do with the data they collect, or worse still, undermine the researcher/participant relationship[83].

Andreotta *et al*. compared the notion of informed consent and how it has been understood and operationalised in the ethical regulation of biomedical research with current AI big data practices[84]. In doing so they recognise that the development phase of a commercial AI system may not be characterised as research or as falling within research regulation, including the need for informed consent[85]. They propose 'soft governance' for commercial uses of big data, where REC-like bodies uphold consent and privacy, but recognise private industry may lack the motivation or culture to adhere to such voluntary oversight.

Solutions have been proposed to improve individual control and choice in how data are used, which may be used individually or in combination:

- **electronic informed consent**[86], in which online forms and communication are used to give consent for various uses of health data
- **dynamic consent**[87], which allows participants to modify their consent periodically for uses that they wish to permit and those that they specifically exclude. Recognising the potential for participant fatigue in this approach, the use of algorithms to predict and mediate dynamic consent decisions for social media data in health research has been explored[88].
- **broad consent**[89], which has been used by large data platforms and biobank to allow individuals to consent to a broad use (e.g. 'health-related research') and in some cases can involve consent to a broad group of potential data users (e.g. including commercial).

The concept of social licence has also been explored as a guideline for ethical governance of health research projects which rely on the use of health data[90]. The concept has been described as referring to "the informal permissions granted to institutions… by members of the public to carry out a particular set of activities"[91]. Social licence focuses on trustworthiness in the stewardship of data uses and can be strengthened by community engagement that aims to recognise the interests and perspectives of the relevant stakeholders.

Utility and value to communities

A key value that is critical for AI health research is community beneficence. Research benefits should be a 'two-way street' meaning that communities that provide data also benefit in some ways from sharing data and the research has utility and value to them. More broadly, data should be shared for

[83] Pickering, B. (2021) Trust, but verify: Informed consent, AI technologies, and public health emergencies. *Future Internet* 13, 132
[84] Andreotta, A.J., Kirkham, N. & Rizzi, M. (2021) AI, big data, and the future of consent. *AI & Soc.*
[85] Referring back to the partnership between Google DeepMind and the Royal Free London NHS Foundation Trust that involved sharing identifiable data from 1.6 million patients. See: Powles, J. & Hodson, H. (2017) Google DeepMind and healthcare in an age of algorithms. *Health Technol.* 7:351–367 and Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)
[86] SageBionetworks (2020) Elements of informed consent toolkit
[87] Kaye J., Whitley E.A., Lund D., et al. (2014) Dynamic consent: A patient interface for twentyfirst century research networks. *European Journal of Human Genetics* 23(2):141–146
[88] Norval C. & Henderson T. (2020) Automating dynamic consent decisions for the processing of social media data in health research. *J Empir Res Hum Res Ethics* 15(3):187-201
[89] Mikkelsen, R.B., Gjerris, M., Waldemar, G. *et al.* (2019) Broad consent for biobanks is best – provided it is also deep. *BMC Med Ethics* 20,71
[90] Muller, S.H.A., Kalkman, S., van Thiel, G.J.M.W. *et al. (2021)* The social licence for data-intensive health research: towards co-creation, public value and trust. *BMC Med Ethics* 22, 110
[91] Shaw, J.A., Sethi, N. & Cassel, C.K. (2020) Social license for the use of big data in the COVID-19 era. *npj Digit. Med.* 3, 128

research where this adds value, not simply sharing for its own sake[92]. In the same vein, it's important to consider when an AI system is the appropriate solution to a problem, and when it is not. Alami *et al.* highlight that some LMICs face the challenge of having to implement and coordinate health care delivered by, or overseen by, international development agencies and NGOs who implement AI-based health applications in silos based on their particular health focus (malaria, maternal health etc.). This runs the risk of paying little attention to other urgent problems and medicalizing certain problems that may be more effectively addressed through poverty reduction, health education, promotion and prevention programs[93].

Partnerships and engagement with multiple stakeholders are required in order to identify and understand local health priorities and potential solutions for communities, and to ensure research is contextualised, culturally grounded and useful. Government agencies, policy makers, academic researchers, local health professionals, community-based organisations, NGOs and industry should be involved in the development and implementation of AI technologies to help maximise local relevance and social value. Alami *et al*. argue that "women, minorities, and poor communities must also play a significant role and have a genuine, legitimate seat at the table in order to guarantee that innovation is truly beneficial, while ensuring that biases and structural inequalities are mitigated"[94].

Reliability and safety

Research on an AI technology destined for use within the health sector can help determine its relevance, validity and reliability. In addition, disclosure by developers regarding the conditions over which an AI system is valid, and disclosure of possible applications or situations where a given system should not be used can promote the reliability of the technology[95]. However, as Nebeker *et al.* point out, not all AI technologies undergo rigorous testing because of how they are developed, by whom and where they are positioned in the regulatory landscape (see Box 2)[96].

Box 2: Case study Moodflex

MoodFlex developed a mobile app to detect signals of poor mental health by analysing a person's typing and voice patterns from their smartphones. The company is negotiating with several municipalities to integrate their product within the public mental healthcare system, but since MoodFlex does not claim to provide a clinical diagnosis or treatment, approval from the US Food and Drug Administration is not necessary. While the vendor claims to have a proven product there are no publications documenting evidence that it is safe, valid or reliable. The only research that is formally acknowledged involves an evaluation of the implementation process and uptake of the product by health providers within the state mental health system. The patient will be invited to download the app after reviewing the vendor's Terms & Conditions – no other consent process is proposed. The algorithm is proprietary, and therefore, an external body is unable to determine whether the algorithm that resulted from a machine-learning process was trained on representative data, or how decision-making occurs. Data captured about people using the app are owned by the vendor.

Extract taken from Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17, 137 (2019). See the paper for a full analysis of the case study.

[92] Global Forum on Bioethics in Research (2019) Meeting report: Ethics of data sharing and biobanking in health research
[93] Alami, H., Rivard, L., Lehoux, P. *et al.* (2020) Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Global Health* 16
[94] *ibid*
[95] Fletcher, R.R., Nakeshimana, A. & Olubeko, O. (2021) Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front. Artif. Intell.* 3
[96] Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17, 137

A lack of research and no formal assessment of reliability may have safety implications, especially where AI is used to deliver treatment, or make decisions in healthcare. In their briefing paper on AI, the Nuffield Council on Bioethics cite a 2015 clinical trial, when an AI app was used to predict which patients were likely to develop complications following pneumonia, and therefore should be hospitalised. The app erroneously instructed doctors to send home patients with asthma due to its inability to take contextual information into account[97]. In another example, internal IBM documents showed that its Watson supercomputer often gave unsafe and incorrect cancer treatment recommendations, a problem largely blamed on the nature of the training data[98]. These issues highlight the importance of research during implementation, involving the continuous, systematic and transparent assessment of AI technology during actual use [99].

Reproducibility

Testing reproducibility is an important aspect of the research process, to validate the research findings and promote open and accessible research[100]. However, how and whether research is carried out to assess an AI application's reproducibility and effectiveness is variable in terms of standards and methods[101]. Kapoor *et al.* cite the reasons for caution concerning the reproducibility of machine learning including that performance evaluation is typically difficult in machine learning and code tends to be complex and as yet lacks standardization[102].

In response to a consultation on the US's Update of the National AI Research and Development Strategic Plan, Kapoor[103] *et al*. recommend that government funding should be conditional on disclosing research materials, such as the code and data, that would be necessary to replicate a study and governments could prioritise funding for setting standards and making tools available to independent researchers to validate claims of effectiveness of AI applications**.** However, issues remain with reproducibility of research by private companies being largely inaccessible due to issues with data sharing and lack of access to computational infrastructure.

Conflict of interest

Some have highlighted the potential conflict of interest in relation to academic and government collaboration with for-profit technology companies[104]. For example, clinicians who contribute patient-derived data sets for training or evaluating AI systems who want to be rewarded or recognised for their contribution to the intellectual property of the AI system that was founded on their diagnostic work and patient medical records[105].

The New York Times reported on Paige.AI, a for-profit start-up founded by senior colleagues at the Memorial Sloan Kettering Cancer Center in Manhattan. A few leading researchers and Center Board

[97] Caruana R, *et al*. (2015) Intelligible models for healthcare, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp1721-30 quoted in Nuffield Council on Bioethics in Research (2018) Bioethics Briefing Note: Artificial intelligence (AI) in healthcare and research

[98] Ross, C. & Swetlitz, I. (2018) IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *STAT News*

[99] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance

[100] Pineau, J., Vincent-Lamarre, P., Sinha, K. *et al*. (2020) Improving reproducibility in machine learning research (A Report from the NeurIPS 2019 reproducibility program)

[101] Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17**,** 137

[102] Princeton University's Center for Information Technology Policy. Project: Irreproducibility in machine learning

[103] Princeton University's Center for Information Technology Policy (2022) Response to request for information to the update of the national artificial intelligence research and development strategic plan

[104] Ochigame, R. (2019) The invention of "ethical AI": How big tech manipulates academia to avoid regulation. *The Intercept*

[105] Abràmoff M.D., Tobey D., & Char D.S. (2020) Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process. *Am J Ophthalmol* 214:134-142

members have a financial stake in the company, which was given exclusive use of the cancer centres vast archive of tissue slides, without the proposal going out for competitive bidding. While start-ups in the biomedical field are not unusual, as the Times points out, what was being commercialised in this case was not a product *per se* but access to raw materials gathered over decades[106].

A question has been raised about whether patients should be told about potential conflicts of interest that are inherent to the AI technology. For example, if the AI technology providing the treatment recommendation was funded and created by the same company whose drugs were prescribed, or if their doctor was involved in the research and development[107].

4. Transparency and engagement

Engagement

Many guidelines call for inclusive and participatory discourse and engagement during the development of AI applications for health as a way to mitigate bias, ensure the benefits of AI are shared widely, to increase citizens' and health care professionals' understanding and trust in the technology and to understand ethical concerns of those involved in the AI lifecycle[108]. But there are currently no widely used and accepted mechanisms or schemes for achieving the input of health professionals, patients or the community in AI development and deployment.

However, Ada Lovelace Institute has proposed a 'framework for participatory data stewardship' in the field of AI which promotes practices that empower people to help inform, shape and in some cases govern their own data. As they explain, the participatory approaches are not proposed "as an alternative to legal and rights-based approaches, but rather as a set of complementary mechanisms to ensure public confidence and trust in appropriate uses of data, and – in some cases – to help shape the future of rights-based approaches, governance and regulation"[109]. Any approach to engagement must be coupled with activities that increase knowledge, understanding, awareness and agency of those being engaged.

Assessments and codes of practice to promote transparency and accountability

The complexity of AI systems calls for transparency and an assessment of impact and risk to help foster trust. A range of AI assessments and codes have been published[110] and proposed in the literature, many of which embed engagement into the process e.g.:

- **Equity assessments**[111]
- **Human rights impact assessments** (HRIA) to identify and mitigate the risks to the rights of people affected by data processing and use of AI in health care [112]. Such an assessment could

[106] New York Times (2018) Sloane Kettering's cozy deal with start up ignites a new uproar
[107] Brummel, E.S. (2017) Confronting natural conflicts of interest and artificial intelligence *Journal of Law and the Biosciences* 4(2)
[108] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)
[109] Ada Lovelace Institute (2021) Report: Participatory data stewardship: A framework for involving people in the use of data
[110] For a range of AI assessments, diagnostics and audit tools, see this spreadsheet (author unknown)
[111] Ferryman, K. (2020) Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. *Journal of the American Medical Informatics Association* 27(12)
[112] For more on HRIAs see: United Nations Special Rapporteur on the promotion and protection of the rights to freedom of opinion and expression A/73/348 (2018) Report on artificial intelligence technologies and implications for freedom of expression and the information environment; European Center for Not-for-Profit Law and Data & Society (2021) Mandating human rights impact assessments in the AI Act; Adams, R., Pienaar, G., Olorunju N. *et al.* (2021) Human rights and the fourth industrial revolution in South Africa; Mantelero, A. & Esposito M.S. (2021) An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review* 41

include the risks to the growing workforce employed to label data and train algorithms, who are often based in low resource settings[113].

- **Human rights, ethical and social impact assessments** as an integrated approach to risk assessment that focuses on human rights and encompasses contextual social and ethical values[114].
- **Environmental impact assessments**. Initiatives are underway to investigate the environmental costs of AI and to find ways of making it more sustainable[115, 116]. This recognises in particular the ethical issue of AI's contribution to climate change[117].
- **NHSX Code of Conduct for Data-Driven Health and Care Technologies** which advocates for the involvement of users of the proposed technology in the development phase[118].

The differently scoped assessment frameworks bring different sets of organisations into formalised relationships with each other, have economic and political consequences and set the conditions for different types of accountability[119].

González-Esteban *et al.* address the issue of accountability more broadly. They propose that in addition to documented procedures for risk assessment and mitigation, a system is needed to ensure the different stakeholders can report concerns and that all AI systems and their development process should be auditable by independent third parties who can look at what was done, and why[120].

Algorithm impact assessments

Algorithm impact assessments (AIA) are an emerging form of AI assessments and have been proposed – or introduced – by governments as part of the governance approach for AI systems, including in the US[121], the EU[122] and Canada[123]. An AIA framework for public agencies has been published[124] and various types of AIAs are also being tested in the private sector[125].

In their policy brief on the draft EU AI Act, the UK's Ada Lovelace Institute points to the need for a sectoral approach to AIA given AI's broad field of application. Of note, the Institute has set out a detailed proposal for an AIA for data access in a healthcare context[126] as part of its broader work looking at the current state of research and practice around algorithmic audits and impact assessments[127]. The assessment is being tested in the context of the UK's NHS AI Lab's National Medical Imaging Platform, a proposed large-scale dataset of high-quality chest X-rays, MRIs, skin, ophthalmology and other images, made available to researchers and private companies to test, train

[113] Financial Times (2019) AI's new workforce: the data-labelling industry spreads globally
[114] University of Oslo's Faculty of Law seminar (2022): Mantelero, A. Towards a new legal framework for AI: Human rights, ethical and social impact assessment in AI
[115] The Sustainable AI Lab website
[116] Mulligan, C., Elaluf-Calderwood, S. (2021) AI ethics: A framework for measuring embodied carbon in AI systems. *AI Ethics*
[117] Nordgren, A. (2022) Artificial intelligence and climate change: ethical issues. *J. Inf. Commun. Ethics Soc.*
[118] UK Government Department of Health & Social Care (2021) A guide to good practice for digital and data-driven health technologies
[119] Watkins, E.A., Moss, E., Metcalf, J. *et al*. (2021) Governing algorithmic systems with impact assessments: Six observations. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21) May 19–21, 2021, Virtual Event, USA. ACM, New York, NY, USA
[120] González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2)
[121] US Algorithmic Accountability Act of 2019
[122] European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence COM (2021) 206
[123] Government of Canada (2020) Algorithmic impact assessment tool; Government of Canada (2020) Canadian Treasury Board Directive on the use of machine learning for decision-making
[124] Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. (2018) AI NOW Report: Algorithmic impact assessments: A practical framework for public agency accountability
[125] Watkins, E.A., Moss, E., Metcalf, J., *et al*. (2021) Governing algorithmic systems with impact assessments: six observations. AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp1010-1022
[126] Ada Lovelace Institute (2022) Report: Algorithmic impact assessment: A case study in healthcare
[127] Ada Lovelace Institute and DataKind UK (2020) Report: Examining the black box: Tools for assessing algorithmic systems

and validate medical AI products[128]. The AIA involves privacy, ethics and human rights considerations and stakeholder engagement, which are set out in a user guide[129].

The Ada Lovelace Institute's work with the NHS AI Lab is exploring whether risk and impact assessment can only best be delivered by external certification before deployment, or if self-certification before deployment backed by external post-deployment audit, might be sufficient. They have also recommended the exploration of innovative engagement methods as part of AIAs such as standing representative panels[130] or 'citizens juries', as well as conventional representation through civil society[131].

Impact assessments can be helpful but may run the risk of distilling ethics into a process in order to 'comply' rather than approaching ethics as a process of reflection and analysis. By completing the assessment, is the developer's or the researcher's ethical responsibility exhausted? Or is the tool one piece of the ethical deliberation and the accountability for the use of AI in health research?

The AI Now Institute has argued that AIAs are explicitly designed to engage a wide range of individuals, communities, researchers, and policymakers in accountability efforts by involving them through the various stages and due process elements of the AIA. This stands in contrast to standard data protection impact assessments, which may also deal with risks of automated systems used to evaluate people based on their personal data, but are not shared with the public, and have no built-in external researcher review or other individualised due process mechanisms[132]. In this sense, involving a wide range of stakeholders may give AIAs more substance and make them more than a 'tick box' exercise.

Metcalf *et al*. point out that the AIA developed to identify and evaluate impacts will shape what harms are detected. That is, the assessment and what it uncovers is influenced by what is asked and who is asking[133]. They call attention to the possible danger that AIAs may become an abstract exercise, which does not account for the harms AI systems can cause in practice. Metcalf *et al*. note that "regulatory agencies, private companies, affected communities, and independent researchers alone [do not possess] enough insight into the design, operation, and effects of AI systems to be able to evaluate the relationship between their impacts and their actual or potential harms". But they suggested that these entities, together, "form an epistemic community" that could help elucidate a more complete set of harms as measurable impacts and form the basis for meaningful accountability[134].

The advent of AIAs has the potential to spawn a new industry of internal and 3rd party assessors, raising the issue of cost and who this will exclude from the assessments. The Ada Lovelace Institute policy brief on the EU AI Act highlights cost as a potential source of uncertainty for everyone involved in the AI lifecycle (investors, developers and those who deploy the AI system)[135]. This issue will be particularly acute for organisations – in HICs and LMICs - who don't have the financial resources to put AIAs into practice.

---

[128] Ada Lovelace Institute (2021) Project summary: Algorithmic impact assessments in healthcare
[129] Ada Lovelace Institute (2022) Resource: Algorithmic impact assessment: User guide
[130] Standing bodies for engagement can lead to professionalised individuals who become unrepresentative of the general population as a result of the knowledge gained in these roles. This raises questions of representativeness – whom is being represented and what characteristics (socio-demographic, health problems, knowledge of AI etc) are needed to ensure such bodies remain representative?
[131] Ada Lovelace Institute (2022) Report: Expert opinion: Regulating AI in Europe
[132] Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. (2018) AI NOW Report: Algorithmic impact assessments: A practical framework for public agency accountability
[133] Metcalf, J., Moss, E., Watkins, E.A., *et al*. (2021) Algorithmic impact assessments and accountability: The co-construction of impacts. In ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Mar 3–10, 2021, Virtual Event, Canada. ACM, New York, USA
[134] *ibid*
[135] Ada Lovelace Institute (2022) Report: Expert opinion: Regulating AI in Europe

Clear and transparent reporting

Transparency can be promoted by standardised reporting of health research involving AI systems. Campbell *et al*. explain that standardisation is important because it enhances the validity, comparability, and usefulness of research and mitigates the potential for clinical decisions causing patient harm or inequity, either because the results are not valid in general or are not generalisable to the particular patient[136]. Relevant guidelines include:

- **The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)-AI**[137] is a reporting guideline for clinical trial protocols evaluating interventions with an AI component.
- **The Consolidated Standards of Reporting Trials (CONSORT)-AI**[138] is a reporting guideline for clinical trial reports. It recommends that investigators provide a clear description of the AI intervention, including instructions and skills required for use, handling of the input/output data of the AI algorithm, defining the inclusion and exclusion criteria for participants and what type of data was studied, the human-AI interaction, how the AI device was integrated into the trial setting, whether the interface and code can be accessed publicly and results of any error cases analyses.

Campbell *et al*. explain that the guidelines are meant to ensure that the entire end-to-end pathway for the technology is reliable and reproducible when applied to a similar population, rather than the algorithm being the only unit of evaluation during the trial[139].

The importance of clear and transparent reporting has been highlighted by Faes *et al*.. Their systematic review of studies using AI for disease diagnosis using medical imaging identified 20,000 studies but found less than 1% had sufficiently high-quality design and reporting to be included in the meta-analysis. They reflect that whereas the medical community is accustomed to complying with international standards of reporting, this appears to be much less prominent in other fields such as statistics, mathematics, or computational science[140]. Guidelines may not be enough: changes in culture are also required for their adoption across the different disciplines involved in the AI health research landscape.

The Ada Lovelace Institute has drawn attention to the use of 'algorithm registers' in Amsterdam and Helsinki that provide a list of AI systems and algorithms put into use by these cities[141]. The registries use easy to understand language, provide some technical detail, and state which ethical principles were employed to mitigate biases and risk[142]. Although not currently used in health research[143], it would be interesting to consider whether they can be re-purposed for this field.

5. Ethics and regulatory oversight

Governance systems can comprise: strategies, guidelines, ethics principles; legislative frameworks; best practice policies and standard operating procedures (e.g. data management plans); ethics

[136] Campbell, J.P., Lee, A.Y., Abràmoff, M., *et al*. (2020) Reporting guidelines for artificial intelligence in medical research. *Ophthalmology* 127(12)
[137] Rivera, S.C., Liu, X., Chan A., *et al*. (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* 370
[138] Rivera, S.C., Liu, X., Moher, D., *et al*. (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 26: 1364–1374
[139] Campbell, J.P., Lee, A.Y., Abràmoff, M., *et al*. (2020) Reporting guidelines for artificial intelligence in medical research. *Ophthalmology* 127(12)
[140] Faes, L., Liu, X., Wagner, S.K., *et al*. (2020) A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. *Trans. Vis. Sci. Tech.* 9(2):7
[141] Ada Lovelace Institute (2021) Report: Participatory data stewardship: A framework for involving people in the use of data
[142] Modjeska, N. (2020) AI registers: Finally, a tool to increase transparency in AL/ML Medium/Towards Data Science website
[143] Though Helsinki has published a register in the field of health care: City of Helsinki AI register: Health centre chatbot

committee review and accountability mechanisms to improve transparency of AI systems and their impact[144]. Actors involved in the development, testing, deployment and use of AI technologies - and in their governance - include AI system and technology developers, researchers, research participants, healthcare professionals, governments, policy makers, private companies, funders and journal editors[145]. The challenges for governance in this space are that AI systems operate at a level of complexity beyond the comprehension of many of those involve in their governance[146] and the governance requirements vary between the developers, evaluators and users of AI and whether they work in the public or private sector.

Ethics oversight

The classic research ethics review frameworks and processes, and the way they are operationalised and institutionalised, are not fit for purpose and are ill-prepared for health research that uses an AI system. Traditionally, computer science may not be subject to the same regulatory requirements as human subject research, or it may have been considered low risk and tended to fall outside of the ambit of RECs. The scale and potential impact of AI systems, and the large volumes of data they require, call for a reconciliation of the values and methods used in these different research cultures[147].

Nebeker *et al.* propose that as with regulated human subject research, a research plan should be developed for digital health research and reviewed by an external and objective REC that will assess the degree of burden on potential participants, potential risks and benefits, and that individuals have the ability to make an informed choice about their participation. However, they go on to articulate the problem:

> "…our regulatory bodies (e.g., IRB) may not have the experience or knowledge needed to conduct a risk assessment to evaluate the probability or magnitude of potential harms. Technologists and data scientists who are making the tools and training the algorithms may not have received ethics education as part of their formal training, which may lead to a lack of awareness regarding privacy concerns, risks assessment, usability, and societal impact. They may also not be familiar with regulatory requirements to protect research participants. Similarly, the training data used to inform the algorithm development are often not considered to qualify as human subjects research, which – even in a regulated environment – makes a prospective review for safety potentially unavailable[148]."

In this respect there are challenges in relation to:

- **Performing an adequate ethics review** due to the complexity of AI systems and/or lack of necessary skill on RECs to perform the risk/benefit assessment, evaluate the consent model and assess complex algorithms, especially in countries where ethical guidelines and regulatory structures for research are mostly limited to clinical and biomedical research. There is also the question of the "minimum" information about the algorithm that should be presented to the REC for adequate understanding and evaluation.

[144] González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2)
[145] Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17, 137
[146] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)
[147] McCradden, M.D., Anderson J.A., Stephenson E.A. *et al*. (2022) A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics* 22(5):8-22
[148] Nebeker, C., Torous, J. & Bartlett Ellis, R.J. (2019) Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med* 17, 137

- **The ethics oversight process, including the role of RECs**. Questions include whether ethics review should be static or ongoing to address the different parts of the AI lifecycle? Should algorithmic impact assessments form part of the ethics review? How can the necessary skills be incorporated onto a REC and/ or is there a need for dedicated (national?) RECs for health research that uses AI?

McCradden *et al.* have articulated the role of RECs in their Research Ethics Framework for the clinical translation of healthcare machine learning [149]:

- "RECs should require reporting metrics (or justification of why they are not included) in their review process to assess potential risks and limitations. This addition would promote scientific consistency and inform the appropriateness of a model's potential clinical integration.
- RECs play a role in evaluating the proposed consent model and whether a waiver of consent is appropriate, but mindful that 'low-risk' interventions can entail higher mortality in the intervention group.
- RECs should suggest to researchers when they need additional perspectives (e.g., social scientists, representatives from marginalized groups, under-served patient navigators and representatives) to better inform the project design and mitigate potential unforeseen risks.
- RECs should aim for representation within their own board or among protocol reviewers to seek the advice of those with specific knowledge about equity (e.g., health equity researchers, ethicists, social scientists, etc)."

Other initiatives working in this area or who have published guidance of relevance to the ethics review process include:

- The African Observatory on Responsible AI[150] is working with the AI4D innovation labs[151] in Africa to develop a road map for establishing an independent REC for AI research in Africa.
- ReCODE has developed tools to help researchers and RECs evaluate AI technology used in health research[152].
- A paper by Ferretti *et al*. which is not specific to AI but addresses ethics review of big data research, which is applicable to AI[153].

Strategies, guidelines and principles

Reflecting the growing importance and uptake of AI around the world, there has been a proliferation of strategies, ethics guidelines and principles published by private companies, government agencies, academic institutes, the public sector and others[154, 155]. A scoping review published in 2021 found that 38 national and international governing bodies have established or are developing AI strategies[156]. While no two are the same, many of the strategies highlight the importance of AI for

[149] McCradden, M.D., Anderson J.A., Stephenson E.A. *et al*. (2022) A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics* 22(5):8-22
[150] African Observatory on Responsible Artificial Intelligence website
[151] Artificial Intelligence for Development Africa website
[152] ReCODE Health. Research tool: The Digital Health Framework
[153] Ferretti, A., Ienca, M., Sheehan, M. *et al.* (2021) Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics* 22, 51
[154] Jobin, A., Ienca, M. & Vayena, E. (2019) Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence* 1:389-399
[155] Sienna Project website (2018) Research ethics codes and guidelines for AI & robotics; Tambornino, L., Lanzerath, D., Rodrigues, R., Wright, D. *et al*. (2019) Sienna D4.3: Survey of REC approaches and codes for artificial intelligence & robotics
[156] Murphy, K., Di Ruggiero, E., Upshur, R. *et al.* (2021) Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 22(14)

health. An earlier review provided a comprehensive account of 26 country and regional strategies and their scope[157].

A meta-analysis covering 84 documents relating to ethical guidelines and principles, revealed:

> "a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented"[158].

Notably, the analysis found a significant representation of more economically developed countries, with the USA and the UK together accounting for more than a third of all published ethical AI principles, while African and South-American countries were not represented independently from international or supra-national organisations. This indicates the lack of global equity in international debate over ethical AI and the need to open the discussion about the ethical rules of AI up to different cultures and philosophical perspectives[159].

Many of the published guidelines and ethics principles apply to the development and use of AI broadly, for example the UNESCO recommendations[160]. In contrast, WHO focussed specifically on AI for health in its guidance *Ethics & Governance of Artificial Intelligence for Health.* Six consensus principles are provided to ensure AI works to the public benefit of all countries. A set of recommendations are also provided to "ensure the governance of AI for health maximizes the promise of the technology and holds all stakeholders – in the public and private sector – accountable and responsive to the healthcare workers who will rely on these technologies and the communities and individuals whose health will be affected by its use"[161]. The guidance was developed with an international expert groups and other stakeholder input and health research is included in its scope. The comprehensive guidance is recommended reading for anyone interested in this year's GFBR theme.

Self-regulation vs regulation

The proliferation of soft law and the advent of the diffuse field of 'AI ethics' reflects a push for self-regulation, especially by private industry[162]. This supposes that principles and codes of practice are sufficient to ensure ethical AI, while legally mandated regulatory standards and enforcement are widely viewed by industry as stifling innovation[163]. The growing instrumentalisation of ethical language by technology companies has coined the phrase "ethics washing"[164] while recent examples show the flaws of self-regulation[165] and concerns about conflict of interest[166].

---

[157] The review included details of strategies for: Australia, Canada, China, Denmark, EU Commission, Finland, France, Germany, India, Italy, Japan, Kenya, Malaysia, Mexico, New Zealand, Nordic-Baltic Region, Poland, Russia, Singapore, South Korea, Sweden, Taiwan, Tunisia, UAE, United Kingdom, United States. See Dutton, T. (2018) An overview of national AI strategies. *Medium*

[158] Jobin, A., Ienca, M. & Vayena, E. (2019) Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence* 1:389-399

[159] Goffi, E.R. (2021) The importance of cultural diversity in AI ethics. Institut Sapiens website

[160] UNESCO website. Recommendations on the ethics of artificial intelligence

[161] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance

[162] Ochigame, R. (2019) The invention of "ethical AI": How big tech manipulates academia to avoid regulation. *The Intercept*

[163] Brummel, E.S. (2017) Confronting natural conflicts of interest and artificial intelligence. *Journal of Law and the Biosciences* 4(2)

[164] Google's ethics committee – the Advanced Technology External Advisory Council – was dissolved soon after it was set up "after leaked details of its members demonstrated that they included people who were avowedly homophobic, xenophobic and misogynistic" González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2). See also Bietti, E. (2020) From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy (2019). Draft - Final Paper Published in the Proceedings to ACM FAT

[165] Powles, J. & Hodson, H. (2017) Google DeepMind and healthcare in an age of algorithms. *Health Technol* 7:351–367

[166] Brummel, E.S. (2017) Confronting natural conflicts of interest and artificial intelligence. *Journal of Law and the Biosciences* 4(2)

Some technology companies[167] have acknowledged that self-regulation isn't enough and argued that government agencies need greater power. Stahl *et al*. point out that ethics and legislative interventions are complementary and mutually supportive and "the question is thus not so much *whether* legal measures should be used to address ethical and human rights issues in AI, but *what shape* they should take to achieve this aim"[168].

The AI Now Institute propose an expansion in the power of sector-specific agencies to oversee, audit, and monitor AI technologies by domain, recognising that a national AI safety body or general AI standards and certification model will struggle to meet the sectoral expertise requirements needed for nuanced regulation[169]. As an example, the USA US Food and Drug Administration (FDA) has published an action plan that would involve it taking on additional regulatory responsibilities of AI-enabled devices in healthcare. In its proposed regulatory plan, the FDA engaged ethical principles, for example, by highlighting the importance of transparency and engagement[170]. The role of regulatory agencies like the FDA raises an interesting question about the extent to which requirements for market authorisation of a new AI-based health application align (or not) with research ethics regulatory requirements. The use of 'sandboxes' has also been proposed, as a form of regulatory experimentation to assess AI medical devices before their widespread adoption and implementation[171].

Despite the challenges of broad regulation, some initiatives are taking place that are moving from a soft law to hard law approach. For example, the new EU Artificial Intelligence Act[172] will be applied in the future by all member states of the European Union with the aim of ensuring the safe and socially acceptable development and application of "trustworthy" AI[173].

González-Esteban *et al*. explain the limitations of such a regulatory approach:

> "the rapid development of AI … require constant hurried revision of the legal-political framework to ensure it does not become obsolete and futile. The problem is that, no matter how hard we try to narrow the gap between detecting changes and their impacts and revising the legislative framework, there is always a time lag, with consequences that are difficult to control. Meanwhile, the fact that legislative frameworks are limited to particular countries limits their results and effectiveness in the hyperglobalised processes of digital transformation. As long as there is no international legislation, the possibilities of regulating the design and use of AI are greatly restricted."[174]

The UK's Ada Lovelace Institute[175] and others[176, 177] have provided a critique of the AI Act to promote debates amongst EU and global policymakers on the shortcomings of the Act as its potential as a global model. Issues include the failure to account for those who are affected by the deployment of

---

[167] Temple, J. (2019) Tech companies must anticipate the looming risks as AI gets creative. *MIT Technology Review*

[168] Stahl, B.C., Rodrigues, R., Santiago, N. & Macnish, K. (2022) A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review* 45

[169] Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. (2018) AI NOW Report: Algorithmic impact assessments: A practical framework for public agency accountability

[170] US Food & Drug Administration (2021) Artificial intelligence/machine learning-based software as a medical device action plan

[171] Leckenby, E., Dawoud, D., Bouvy, J. *et al.* (2021) The sandbox approach and its potential for use in health technology assessment: A literature review. *Appl Health Econ Health Policy* 19, 857–869

[172] European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence COM (2021) 206

[173] The draft AI Act "would apply to both public and private actors, providers and users of such systems, irrespective of whether they are established within the Union or in a third country" Lilkov, D. (2021) Regulating artificial intelligence in the EU: A risky game. *European View* 20(2):166-174

[174] González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2)

[175] Ada Lovelace Institute (2022) Policy Briefing: People, risk and the unique requirements of AI

[176] European Center for Not-for-Profit Law and Data & Society (2021) Mandating human rights impact assessments in the AI Act

[177] González-Esteban, E. & Calvo, P. (2022) Ethically governing artificial intelligence in the field of scientific research and innovation. *Heliyon* 8(2)

an AI system – in relation to their right to information and to bring complaints; the need to establish clear, judicially reviewable criteria for placing AI systems into categories of risk; extending the meaning of risk to include systemic and environmental risks; and the recommendation for periodic post-deployment assessments in order to evaluate real-world impacts of high-risk AI[178].

In addition to the AI Act other regulatory development in the EU are likely to influence the development of AI for health. For example, the proposed regulation for the European Health Data Space (EHDS) that was released on May 2022[179]. The intention of the EHDS is to increase health data sharing and access all in light of needs for data sets that will be used to train AI models. One potential challenge for the regulatory movement in Europe is to ensure coherence amongst the several instruments that impact AI development and use (e.g. Medical Devices Regulation, *In Vitro* Diagnostic Medical Devices Regulation, the General Data Protection Regulation, AI Act, EHDS regulation). The movement also presents a challenge globally, in that non-EU countries will have to abide by the EU rules if they want to collaborate with EU partners, even if the rules are not suitable for their context. This may be particularly problematic for LMICs and compound existing disparities of power.

6. Additional governance considerations

Human rights approach

Some have suggested a human rights approach (HRA) as providing normative and legal guidance for those developing AI regardless of country or jurisdiction as well as a shared normative language[180, 181]. Wong explains that "the appeal to human rights in HRA serves two purposes, i.e. (i) it offers a normative standard to identify, anticipate, and evaluate the harmful (or beneficial) impacts of AI technologies and (ii) a set of legal and institutional measures to prevent, mitigate, and rectify the harm caused by them based on the existing legal and institutional frameworks". However, they go on to question the universality of the approach given the contested nature and interpretation of human rights, the power asymmetry in shaping the human rights agenda and the need for greater cultural diversity in the human rights debate[182].

Yeung *et al.* have outlined a comprehensive rationale for a 'human rights-centred design, deliberation and oversight' governance framework. The framework is "(1) anchored in human rights norms and a human rights approach (2) utilises a coherent and integrated suite of technical, organisational and evaluation tools and techniques, that is (3) subject to legally mandated external oversight by an independent regulator with appropriate investigatory and enforcement powers, and (4) provides opportunities for meaningful stakeholder and public consultation and deliberation." Recognising that more theoretical and applied research is required to flesh out the details of their proposed approach, an agenda for further research is proposed[183].

Responsibility and liability

The issue of responsibility and liability for errors in the application of AI technology are complex and will likely need to be decided based on the context[184]. Responsibility and liability in research and for

[178] Ada Lovelace Institute (2022) Policy Briefing: People, risk and the unique requirements of AI
[179] European Commission (2022) Press release: European Health Union: A European Health Data Space for people and science
[180] Latonero, M. (2018) Data&Society report: Governing artificial intelligence: Upholding human rights & dignity
[181] Article 19 (2019) Report: Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence
[182] Wong, P.-H. (2020) Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy & Technology* 33:705-715
[183] Yeung, K., Howes, A. & Pogrebna, G. (2019) AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing. Forthcoming in M Dubber and F Pasquale (eds.) The Oxford Handbook of AI Ethics, Oxford University Press (2019)
[184] Jha, S. (2020) Can you sue an algorithm for malpractice? It depends. *STAT News*

clinical and health care decisions may be more defined or diffuse depending on whether the system was privately designed, or a system developed in partnership with a healthcare institute that was trained using data from its patient population and intended for use in the patient population[185]. The Ada Lovelace Institute suggests a "nuanced appraisal must be made of what duties should lie where at what point in time, and who is empowered either legally or by practical control, power or access to data and models, to make changes". They propose that responsibility should not be allocated to the deployer alone, since the power to control and modify such infrastructure, alongside technical resources, largely lies with the AI system developer/provider[186].

WHO has pointed out that liability rules play an important role in promoting safety and accountability, and in some contexts may be the first and only line of defence against errors made by AI technologies. They note that liability regimes for AI technologies are mostly developing in the EU and the USA, with many LMIC yet to adopt an approach. While still lacking sufficient regulatory capacity to assess drugs, vaccines and devices, LMICs may struggle to accurately assess and regulate AI technologies[187].

16 May 2022

[185] Abràmoff M.D., Tobey D., & Char D.S. (2020) Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process. *Am J Ophthalmol* 214:134-142
[186] Ada Lovelace Institute (2022) Report: Expert opinion: Regulating AI in Europe
[187] World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance