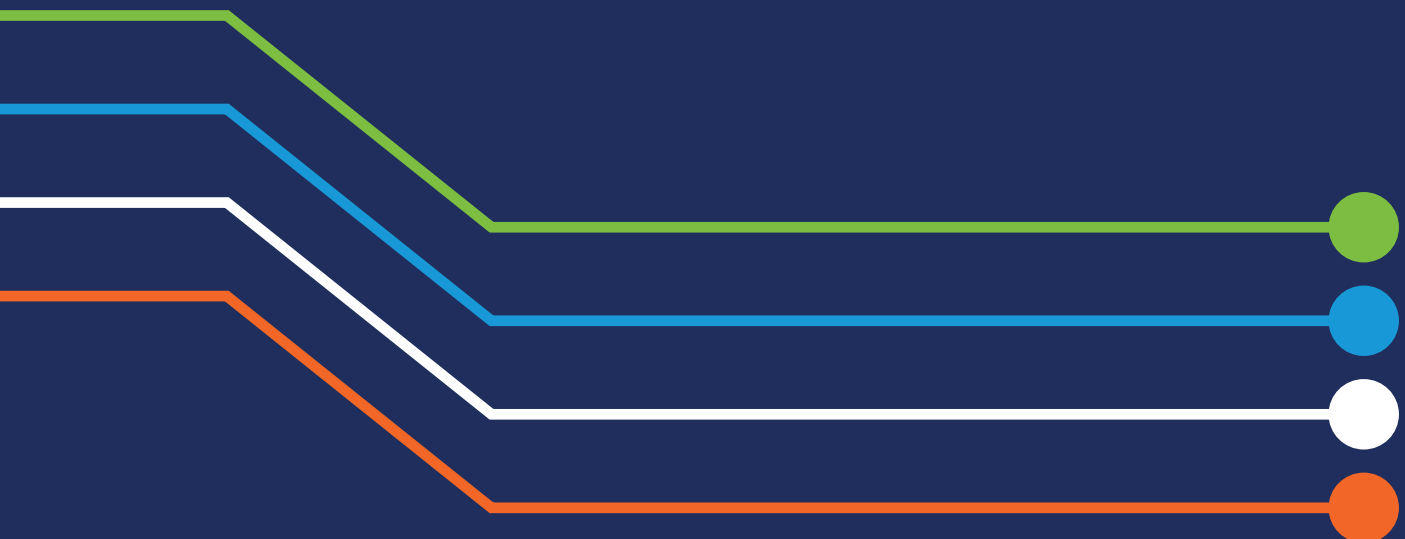




II. Ethics and governance of LMMs in health care and medicine



The ethical principles defined by the WHO expert group (see above) provide guidance to stakeholders on the basic ethical requirements that should direct their decisions and actions in development, deployment and assessment of the use of LMMs in health care and medicine. The principles should be the basis for how governments, public sector agencies, researchers, companies and implementers govern the use of LMMs.

Governance comprises the steering and rule-making functions of governments and other decision-makers, including international health agencies, for achieving national and global health policy conducive to universal health coverage. Governance is also a political process that involves balancing competing influences and demands (1). Current laws and policies are unlikely to suffice for effective management of the use of LMMs, as many were written before the release of earlier versions of LMMs. Governance of LMMs, as with overall governance of AI, involves applying current and new legislation and regulations, “soft law” (such as ethical principles), human rights obligations, codes of practice and internal procedures of companies, industry associations and standard-setting bodies.

Currently, LMMs are deployed more quickly than our ability to fully understand their capability and frailty. An early suggestion for addressing concern about LMMs was for a ban or a moratorium on their development (118). While some countries do restrict the use of or even ban certain LMMs, most governments now seek to ensure that their use can be directed towards socially beneficial outcomes through appropriate governance. Leading AI companies have also called for careful, deliberative development of LMMs and other forms of AI. Neither governments nor companies, however, are immune to competition. Several governments are locked in an “arms race” to technological supremacy, while even AI companies that are calling for regulation are not immune to commercial pressure (119). While optimists consider that many of the challenges and risks of AI can be addressed through design, including ever-larger data sets and more powerful algorithms, critics have pointed out that the limitations of LMMs are systemic and that increasing the size of training data and model parameters will not overcome shortcomings but will in fact amplify them (59).

Governance of LMMs must keep pace with its rapid development and increasing uses and should privilege neither governments seeking a technological advantage nor companies seeking commercial gain. The initial suggestions and recommendations below place ethical principles and human rights obligations at the centre of appropriate governance, comprising both procedures and practices that could be introduced by companies and laws and policies enacted by governments.

LMMs can be considered products of a series (or chain) of decisions on programming and product development by one or more actors. Decisions made at each stage of an AI value chain may have both direct and indirect consequences on those that participate in the development, deployment and use of LMMs downstream. The decisions can be influenced and regulated by governments that enact and enforce laws and policies nationally, regionally and globally. The AI value chain begins with integration of several inputs, which comprise the “AI infrastructure”, such as data, computing power and AI expertise, to the development of general-purpose foundation models. These models can be used directly by a user to perform

various, often unanticipated tasks (including those related to health care). Several general-purpose foundation models are trained specifically for use in health care and medicine.

Appropriate governance of LMMs used in health care and medicine should be defined at each stage of the value chain, from collection of data to deployment of applications in health care. Therefore, the three critical stages of the AI value chain discussed are:

the design and development of general-purpose foundation models (design and development phase);

- definition of a service, application or product with a general-purpose foundation model (provision phase); and
- deployment of a health-care service application or service (deployment phase).

At each stage of the AI value chain, the following questions are asked.

- Which actor (the developer, the provider and/or the deployer) is best placed to address relevant risks? What risks should be addressed in the AI value chain?
- How can the relevant actor(s) address such risks? What ethical principles must they uphold?
- What is the role of a government in addressing risks? What laws, policies or investment might a government introduce or apply to require actors in the AI value chain to uphold specific ethical principles?

During the design and development phase, the focus is on the practices that developers can introduce to uphold ethical commitments and norms and government policies and investments. During the provision phase, the focus is on the measures that governments can introduce to assess and regulate use of LMMs in health care and medicine. During the deployment phase, measures are used by governments and all actors in the value chain to ensure that any potential or actual harm to users is identified and avoided.

4 Design and development of general-purpose foundation models (LMMs)

General-purpose foundation models are usually trained on a vast amount of data, requiring tremendous computing power. Development of LMMs also requires specialized human resources, including scientific and engineering expertise. The WHO guidance on ethics and governance of AI for health (1) recommends that developers of medical AI “should invest in measures to improve the design, oversight, reliability and self-regulation of their products”.

Although most findings and recommendations below could apply to all general-purpose foundation models, the guidance is intended for such models that may be or are used in health care and medicine (either directly by a user or through an application or service). The recommendations below are also intended to guide the design and use of LMMs trained specifically for use in health care and medicine, which may be used directly by users or through an application or service.

4.1 Risks to be addressed during the development of general-purpose foundation models (LMMs)

The design and development of general-purpose foundation models can introduce serious risks that, if left uncorrected, could have either a broad societal impact or specific negative consequences on the users of an LMM. Elimination or mitigation of such risks is the responsibility of the developer, because it is only the developer who can (or could) make certain decisions during design and development, which are beyond of the control of providers and deployers that may use the algorithm (and which cannot be mitigated by correct use of the technology by a provider, deployer or user) (120). The decisions refer, for example, to the data used to train an LMM (121). Obligations to ensure data protection and quality and to mitigate bias are also outside the control of the downstream developer of an application (121), as are measures that might have to be introduced to ensure that LMMs do not issue “AI-fuelled toxicity” (122). Failure to hold the developers of LMMs accountable for such design flaws shields the companies with the most resources from, as one report noted, “responsibility to tackle problems...which their methods may be thoughtlessly baking in as they rush to dominate a new form of applied AI” (122).

At least eight risks should be addressed by the developer of a general-purpose foundation model, including through government laws and regulations:

- bias (associated with the design and training data);

- privacy (of training and other input data);
- labour concerns (outsourced filtering of data to remove offensive content);
- the carbon and water footprints;
- false information, hate speech or misinformation;
- safety and cybersecurity; and
- preserving the epistemic authority of humans
- exclusive control of LMMs

4.2 Measures developers can take to address risks with general-purpose foundation models (LMMs)

A developer could use many measures or practices to address such risks, whether as a commitment to ethical principles or policies or to meet the requirements of governments.

AI expertise (scientific and engineering personnel): A developer can ensure that its scientific and programming personnel can identify and avoid risks. The WHO ethics guidance (1) made several recommendations for the training of scientific and engineering personnel and on the inclusiveness of the design process. In particular, the WHO expert group recommended that developers consider “licensing or certification requirements for developers of ‘high-risk’ AI, including AI for health”.

Companies and other entities that develop an LMM that shall or could be used in health care, scientific research or medicine should consider certification or training to align themselves with requirements in the medical profession and also to increase trust in their products and services (1). Any standards, introduced and enforced by either developers or professional societies, should be written in collaboration with or by government regulators and should be consistent with the WHO ethical principle of promoting human well-being, safety and the public interest. Developers that do not intend but can foresee that their LMM might be used in health may wish to ensure internal expertise to anticipate and address such uses.

Data: While human resources and computing power are essential for the development of LMMs, data are probably the most critical infrastructure requirement. The quality and type of data used to train LMMs determine whether it meets core ethical principles and legal requirements (123). Although AI developers have agreed in qualitative surveys that data quality “matters” and require a significant commitment of time, work related to data is often under-valued, which can have significant negative repercussions for AI in “high-stakes” domains such as health care and medicine (123). If data are not of the appropriate quality, several WHO guiding principles could be violated, including the promotion of human well-being, safety and the public interest and the principle of ensuring inclusiveness and equity if the data are biased.

As use of data for health care will probably require strict adherence to laws for informed consent, developers who train LMMs intended for use in health care and medicine might have to rely on smaller data sets (59). Smaller data sets might also be preferable to ensure the quality of data, that the data are diversified in order to avoid bias (59) and that they reflect the composition and reality of the population(s) that will be served by the LMM. Smaller data sets might, however, increase the risk of re-identification of individuals, which would expose them to current or future harm. Reliance on smaller data sets may have further benefits, including reducing the carbon and water footprints (112) of the models and also making it possible for smaller entities to participate in or develop LMMs that require fewer data, computing, human and financial resources (59). Regardless of the size of the data set, developers should undertake “data protection impact assessments”, which would, as required under the General Data Protection Regulation, require developers to assess the risks of data-processing operations to the rights and freedom of individuals and their impact on the protection of personal data (1) before processing such data. Collection of data from low- and middle-income countries could amount to “data colonialism”, in which data are used for commercial or non-commercial purposes without due respect for consent, privacy or autonomy (1).

The assessments could extend beyond risks to privacy and include the quality of data, such as whether they are unbiased and accurate. AI researchers who examine or audit such datasets are reluctant to invest their resources, noting that, while creation of a data set for AI is simple, auditing is difficult, time-consuming and costly. As one researcher noted: “Doing the dirty work is just a lot harder” (124).

Developers can take other measures to improve data quality and adhere to data protection laws. Irrespective of the model size, developers should, in contrast to how early LMMs were developed, train LMMs on data collected according to best-practice data protection rules. Developers should thus avoid using data from third-party sources such as data brokers, as their data may be old, biased, combined incorrectly or have other flaws that may not have been corrected (125). Careful collection of data could also ensure that an LMM does not violate copyright or data protection laws, for which there may be legal repercussions, such that certain LMMs could be labelled unlawful (126).

If third-party data providers are used, they could, for example, be certified, in order to build trust and to ensure their expertise and legitimacy (127). All data used to train LMMs by developers, whether collected directly or from third parties, must be kept up to date. As noted above, some of the leading AI models were not trained with up-to-date data (38), which can jeopardize the performance of the model in health care and medicine, in which new evidence and information meaningfully affect decisions. Data sets should be updated and accurate so that LMMs are appropriate and relevant for the contexts in which they are used.

It may be difficult to ensure that data are adequately transparent. Companies that launch new LMMs have become more and more opaque about the data used to train its models. One leading AI company that released a new LMM stated that: “Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains

no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar” (128).

Unwillingness to be transparent about data is, however, inconsistent with the WHO ethical principle of ensuring transparency, “explainability” and intelligibility. Developers should be transparent about the data they used to train a model so that downstream users, including those who fine-tune the LMM, use the LMM to develop a health-care application and those who use the LMM directly, are aware of any insufficiency or incompleteness of the training data set.

When developers improve data quality by using data workers in low- and middle-income countries to screen the content for abusive, violent or offensive material and to annotate data, those workers should be paid a living wage and be provided with mental health services and other forms of counselling; and developers should introduce safeguards to protect workers from any distress. Governments should update their labour standards to extend benefits to all data workers, to promote a “level playing field” among companies and to ensure that labour standards are maintained and improved over time.

Ethical design and design for values: One approach to integrating ethics and human rights standards into the development of AI technologies is “design for values”, a paradigm for basing design on the values of human dignity, freedom, equality and solidarity and for considering them non-functional requirements (1). Several recommendations in the original WHO expert guidance for the design of AI technologies, including “design for values”, bear repeating here.

The guidance recommended that the design and development of AI technologies not be done solely by scientists and engineers and that “potential end-users and all direct and indirect stakeholders should be engaged from the early stages of AI development in structured inclusive, transparent design and given opportunities to raise ethical issues, voice concerns and provide input for the AI application under consideration” (1). Thus, in development of foundation models, people who might either use or benefit from the models could be involved in the initial development. One proposal is to introduce so-called “human oversight colleges”, which would facilitate inclusion of patient representatives in the development of an LMM that is intended to benefit a patient or caregiver, either directly or indirectly through a medical provider.⁶ Medical and health-care professionals, research scientists, patients, laypeople and vulnerable populations could also be included in the design of LMMs, labelling of data and testing. Inclusiveness in the design of the LMM could, for example, protect human autonomy, as the participation of medical providers could forestall or reduce automation bias by providers. Inclusive design could promote the WHO guiding principle of ensuring inclusiveness and equity, especially if design teams include diverse viewpoints by age, ability, race, ethnicity, sex or gender identity.

⁶ Communication from David Gruson, WHO expert on ethics and governance of AI for health.

The original WHO guidance also recommended that “designers and other stakeholders should ensure that AI systems are designed to perform well-defined tasks with the accuracy and reliability necessary to improve the capacity of health systems and advance patient interests. Designers and other stakeholders should also be able to predict and understand potential secondary outcomes” (1). Even before initiating development of an LMM, the developer could conduct a so-called “pre-mortem” (33) to consider “hypothetical failures”, so that the development team can reverse-engineer such unanticipated failures. This allows developers to identify known and unknown risks and to formulate alternatives (33). A second suggestion, by several developers of general-purpose foundation models, is “red teaming” (129), an evaluation of a model or system that identified vulnerability in real-world simulations that might result in undesirable behaviour, such as an LMM providing a biased opinion, so that the developer can correct the model or system to ensure its reliability and safety. One company announced that it would submit its latest LMM models to the DEFCON conference, a hacker convention, in August 2023, so that “experts can further analyse and stress test their capabilities” (130).

The original WHO guidance also recommended that “the procedures that designers use to ‘design for values’ should be informed and updated by the consensus principles, best practices (e.g. privacy preserving technologies and techniques), standards of ethics by design and evolving professional norms” (1). Appropriate design could limit unauthorized disclosure of data entered into an LMM or address environmental (carbon and water) concerns associated with the training and use of LMMs (see below). It could also ensure that users know that the content produced by an LMM is generated by an AI system and not a human, in order to avoid displacing humans from the centre of epistemic authority. Such notification can remind users, communities and societies that, while an LMM can produce useful information, it cannot be a substitute for knowledge production by humans.

Design to respect environmental considerations: As discussed above, a major concern with respect to LMMs is their carbon and water footprints. Developers should take all possible steps to reduce energy consumption, such as by improving a model’s energy efficiency, and several large technology companies are experimenting with such approaches. For example, one company developed an LMM that is combined with an external database that operates more efficiently than an LMM, which is trained with more variables and outperforms less energy-efficient LMMs (112). Another company is experimenting with an LMM that is based not on one neural network but distributes its variables among 64 smaller neural networks. It is trained to use only two neural networks to complete each task, thereby using only a small percentage of its variables for making each inference (112).

Another means for improving energy efficiency is to develop smaller LMMs that are trained on smaller data sets and which therefore do not require as much energy to train or to operate. Smaller LMMs may not only reduce energy consumption but also open opportunities for smaller companies or entities to develop LMMs and improve the accuracy of outputs (59). Smaller LMMs might be particularly useful for developing “specialized LMMs”, such as those intended specifically for use in health care, scientific research and medicine. Several such LMMs have been introduced, including some developed by large technology companies (59).

4.3 Government laws, policies and public sector investments

Several existing or potential laws or policies could be enforced or written to reduce or avoid risks during the design and development of general-purpose foundation models. Furthermore, governments could make public-sector investments to promote or support the ethical design and development of general-purpose foundation models.

Laws and policies governing use of data: WHO supports application and enforcement of standards, including data protection rules, that govern how data have and will be used to train LMMs. Data protection laws are usually based on rights-based approaches and include standards for regulation of data-processing that both protect the rights of individuals and establish obligations for public and private data controllers and processors and include sanctions and remedies in case of actions that violate statutory rights. As data protection laws have been adopted in over 150 countries, they provide a solid foundation for the development of all AI technologies, including LMMs (1). A limitation of data protection laws is that most were enacted before the emergence of generative and other types and uses of AI, and data protection authorities may be unwilling to apply them too aggressively, as the original laws may not have had the same intent (120).

One requirement of data protection that should be enforced, especially for health data used to train LMMs, is that the data are obtained and processed lawfully. This will often require provision of meaningful informed consent by a data subject for use of their data for the stated purpose. Any further processing should have its own legal basis, as further processing cannot be assumed to be compatible with the original purpose. Companies and other entities that have developed and released LMMs are already under scrutiny for potential use of data obtained without informed consent. The pursuit of ever-larger LMMs, requiring increasingly larger data sets, may lead developers to ignore legal requirements (83). This would also violate the WHO guiding principle of protecting human autonomy. Thus, the WHO expert group recommended that governments “should have clear data protection laws and regulations for the use of health data and protecting individual rights, including the right to meaningful informed consent”.

Other government measures to oversee and regulate the collection and use of data for training LMMs include regulations for generative AI enacted by the Government of China that came into effect in August 2023. The Cyberspace Administration of China imposes several obligations, including that: (i) providers shall employ effective measures to avoid discrimination and bias in selecting training data, (ii) providers use clear labelling and assess the quality of data labelling; and (iii) developers take “effective measures” to meet the goals of authenticity, accuracy, objectivity and diversity of data (131). The requirements are not expected to be applied strictly to companies, which will be required only to take effective measures to ensure appropriate data quality. The measures will apply only to those companies that offer services to the Chinese public (132).

Legislative provisions related to data could include requirements requirements to describe the data sources used to train a foundation model and to use data that are subject to data governance, including for suitability, bias and appropriate mitigation (133).

Other measures that governments could take during design and development are described below.

- *Target product profiles:* Governments and international agencies could issue target product profiles to state the preferences and characteristics of LMMs intended for use in health care and medicine, especially if governments anticipate purchasing such technologies for use in government-run health systems.
- *Design and development standards and requirements:* Governments could require developers to ensure that the design and development of a general-purpose foundation model achieve certain outcomes throughout its life cycle. They could include requirements for the predictability of the model and its interpretability, corrigibility, safety and cybersecurity (134).
- *Pre-certification programmes:* Regulatory agencies could introduce legal obligations and establish incentives to both require and encourage developers to identify and avoid ethical risks, such as bias or undermining autonomy, through measures including pre-certification programmes (1). The previous WHO AI ethics guidance recommends that “government regulators should provide incentives to developers to identify, monitor and address relevant safety- and human rights-related concerns during product design and development and should integrate relevant guidelines into precertification programs” (1).
- *Audits:* Governments could introduce audits of the initial stages of development of foundation models. One proposal is for three types of audit: a “governance audit” of an LMM provider, an audit of the LMMs and an “application audit” of downstream products and services built on LMMs, which would not apply during development of an LMM (121). Audits could be integrated into requirements for approval of LMMs intended for use in health care or medicine (see below). In order for audits to be effective, their quality should be assessed to ensure that they fulfil their intended purpose.
- *Environmental footprint:* Governments could require developers of general-purpose foundation models to address concern about their carbon and water footprints. For example, governments could require developers to measure their energy consumption, to reduce energy use during training (133) and to meet as-yet undefined environmental standards (134).
- *Notification that content from an LMM is “machine-generated”:* Governments could require developers to ensure that any deployment of a general-purpose foundation model includes notification and reminders to end-users that the content was generated by a machine and not a human being (133).

- Governments might also consider requiring or creating incentives for developers to register early-stage AI algorithms or systems to be used in health care and medicine. Early registration could encourage publication of negative results, prevent publication bias or over-optimistic interpretation of results and facilitate integration of knowledge that benefits patients.

Public infrastructure to develop LMMs in the public interest: As uses of LMMs for health proliferate, development of LMMs that adhere to ethical principles could be encouraged by the provision of not-for-profit or public infrastructure, including computing power and public datasets. Such infrastructure, which could be accessible to developers in the public, private and not-for-profit sectors, could require users to adhere to ethical principles and values in exchange for access. It could also help to avoid exclusive control of an LMM by a developer and "level the playing field" between the largest companies and developers that do not have access to such infrastructure and resources.

Governments, subject to independent oversight, could construct infrastructure that is then used by developers to construct LMMs for health care and medicine. For example, an international team of 1000 academic volunteers, the company Hugging Face, and others, with funding from the French Government, trained an LMM called BLOOM with 175 billion parameters, which also required US\$ 7 million of computing time (112).

Efforts to level the playing field are also applicable to academia and its resource disadvantage. The Canadian Government's national Advanced Research Computing Platform serves the country's academic sector, the Chinese Government has approved a national computing power network system to enable academics and others to access data and computing power, and, in the USA, the National AI Research Resource task force has "proposed the creation of a public research cloud and public datasets" (101). There have also been calls by civil society in Europe for governments to play a more assertive role in build so-called "European Large Generative Models", for which AI-specific computing, data infrastructure, science and research support would be provided by governments (135).

4.4 Open-source LMMs

The role of open-source LMMs in integrating ethical principles and addressing known risks is uncertain. Generally, transparency and participation can be increased by using open-source software for the design of an AI technology or by making the source code of software publicly available (1). Open-source software is open to both contributions and feedback, which allows users to understand how the system works, to identify potential issues and to extend and adapt the software (1). Open-source LMMs may present an opportunity to address some concerns about use of LMMs in health care. As open-source models are neither proprietary nor closed, they allow smaller firms and entities, such as not-for-profit institutions, to design LMMs at lower cost (136). LMMs built on open-source models can be scrutinized, as the code and data are available for review. Engagement and policing by a community of users helps to ensure the robustness of open-source models in the long term (136).

Open-source LMMs may not endure, however, if large technology companies that previously made their models available choose not to continue to do so (10). Development of most open-source LMMs was based on an LMM released on a limited basis by Meta (formerly Facebook) (10). Since the LMM and its weights were leaked (137), the company has stated its commitment to open-source approaches, noting that openness “leads to better products, faster innovation, and a flourishing market, which benefits [Meta] as it does many others... ultimately, openness is the best antidote to the fears surrounding AI” (130). Independent observers have noted, however, that, while Meta has made its LMM available on a non-commercial basis, its terms of use includes restrictions, and it is thus not providing its LMM in a manner consistent with open-source principles (138,139).

Additional requirements for use of open-source models in order to monitor their performance and outcomes will be difficult for developers to address; however, the benefit of such models cannot supplant the necessity for regulation and avoidance of harm, such as security concerns associated with use of open-source models (140). Open-source models are vulnerable to misuse (141) and can be attacked to exploit such vulnerability (142). A group of researchers recently found that methods tested on open-source AI systems circumvented AI safety measures and safeguards and could also bypass the safeguards of so-called closed systems (143). Ultimately, open-source models are based on the same black-box technologies used in other LMMs.

One way of encouraging open-source LMMs would be for governments to require that foundation models built with government funding or intellectual property be widely accessible, in the same way that governments have required open access to government-funded research. Governments could also encourage open-source research and development in public facilities, including next-generation models, under controlled conditions with public oversight. Public oversight and participation might be better than the new reality in which Meta’s leaked model allows anyone to “download it and run it on a MacBook M2” (144).

Recommendations:

- Developers that design an LMM that shall or could be used in health care, scientific research or medicine shall consider ethics certification or training for programmers. This would bring AI developers in line with requirements in the medical profession and increase trust in their products and services.
- Regardless of the size of the dataset, developers should undertake “data protection impact assessments” before processing such data, which would require developers to assess the risk that data processing operations would go against the rights and freedom of individuals and its impact on the protection of personal data.
- Developers should train LMMs on data collected according to best-practice data protection rules.

- All data sets used to train LMMs, whether collected directly or via third parties by developers, should be kept up to date and appropriate for the contexts in which the system may be used.
- Developers should be transparent about the data used to train a model, so that users, including those who fine-tune the LMM, use the LMM to develop a health-care application or use the LMM directly, are aware of any insufficiency or incompleteness of the training data set.
- Developers should pay data workers a living wage and provide them with mental health services and other forms of counselling. Developers should also introduce safeguards to protect workers from any distress. Governments should update labour standards to extend such benefits to all data workers, to promote a “level playing field” among companies and to ensure that such labour standards are maintained and improved over time.
- Developers should ensure that LMMs are designed not only by scientists and engineers. Potential users and all direct and indirect stakeholders, including medical providers, scientific researchers, health-care professionals and patients, should be engaged from the early stages of AI development in structured, inclusive, transparent design and given opportunities to raise ethical issues, voice concerns and provide input for the AI application under consideration. Such input could be provided through “human oversight colleges”.
- Developers should ensure that LMMs are designed to perform well-defined tasks with the necessary accuracy and reliability to improve the capacity of health systems and advance patient interests. Developers should also be able to predict and understand potential secondary outcomes. Techniques to meet such requirements include “pre-mortems” and “red teaming”.
- Procedures used by developers to “design for values” should be informed and updated by consensus, best practices (e.g. technologies and techniques to preserve privacy), standards of ethics by design and evolving professional norms, including disclosure that content produced by an LMM is generated by an AI system.
- Developers should take all possible steps to reduce energy consumption (such as by improving the energy efficiency of a model).
- Governments should have strong, enforced data protection laws and regulations for the use of health data that apply to the development of LMMs. The laws must effectively protect people’s rights and give people the tools they need to protect their rights, including the right to meaningful informed consent. Additional tools are likely to be needed for data that are collected and processed for use of LMMs in health care.

- Governments and international agencies such as WHO should issue “target product profiles” to delineate preferences and characteristics of LMMs intended for use in health care and medicine, especially if governments anticipate eventual purchase of such tools for use in government-run health systems.
- Governments should require developers to ensure that the design and development of a general-purpose foundation model will achieve certain outcomes during the product’s life cycle. These could include requirements for the predictability of the model and its interpretability, corrigibility, safety and cybersecurity.
- Regulatory agencies should introduce legal obligations and establish incentives, such as pre-certification programmes, to require and encourage developers to identify and avoid ethical risks, including bias or undermining autonomy.
- Governments should introduce audits of the initial stages of development of foundation models.
- Governments should require developers of general-purpose foundation models to address concerns about the carbon and water footprints of general-purpose foundation models.
- Governments should require developers to ensure that, in any use of a general-purpose foundation model, users are notified and reminded that the content has been generated by a machine and not a human being.
- Governments should consider requiring or creating incentives for developers to register early-stage AI algorithms or systems that are to be used in health care and medicine. Early registration could encourage publication of negative results, prevent publication bias or over-optimistic interpretation of results and could facilitate inclusion of knowledge that benefits patients.
- Governments should invest in or provide not-for-profit or public infrastructure, including computing power and public data sets, accessible to developers in the public, private and not-for-profit sectors, that requires users to adhere to ethical principles and values in exchange for access.
- Governments should encourage the development of open-source LMMs by requiring that foundation models built with government funding or intellectual property are widely accessible, in the same way that governments have required open access to government-funded research. Governments should support open-source research and development in public facilities, including next-generation models, under controlled conditions, with public oversight.

5 Provision with general-purpose foundation models (LMMs)

The uses of general-purpose foundation models depend on whether a user prompts an LMM to generate health-care related outputs or a provider is permitted by the developer to integrate the LMM into a health-care related application, product or service. Either case introduces novel risks that must be addressed by developers, providers or both. Governments are responsible for assessing and regulating uses of such technologies before their deployment.

5.1 Risks to be addressed when providing a health-care service or application with a general-purpose foundation model (LMM)

There is likely to be disagreement about whether both general-purpose foundation models and applications should be assessed and approved when they are used in a product for health-care purposes or used directly by a user. Several of the largest technology companies have been quietly lobbying government officials (for example in the European Union) to abandon an evaluation framework for LMMs and to focus oversight instead on applications that might be used in ways that a government might consider “risky” (145). This would concern both providers that generate and market health-care applications that include the foundation model, and users, such as a provider or patient, who choose to use the LMM directly or indirectly via an AI system. The companies argue that oversight of the general-purpose foundation model would “completely shift the burden” to developers and that others in the value chain should also assume responsibility (145).

While it may not be appropriate to hold the developer of a general-purpose foundation model responsible for all uses of the LMM, it would also be inappropriate to place the burden solely on providers, deployers or users, as they will not have been involved in development of the model and may not understand the associated limitations and risks. This would allow developers of general-purpose foundation models, despite their significant power, resources, oversight and understanding of LMMs, to escape responsibility and would open a “massive hole” in attempts to govern AI technologies for health (145).

A developer may seek to avoid use of an LMM for a health-care purpose (or another use). If a developer does not wish an LMM to be used for health or medical purposes (especially in clinical medicine), it could discourage such use either by preventing entities that develop applications for health or medicine to use (license) the LMM on an application programming interface, or, if the LMM is used directly by a user (provider or patient) for a health-care

purpose, by blocking queries or attaching a clear warning to any responses that include health or medical information and to direct users to information or services that can provide appropriate assistance.

If such measures are not taken or if the developer intends that users apply its LMM for health care, either directly or indirectly through a provider, the developer will have specific responsibilities that only it can satisfy. Furthermore, both developers and providers have further obligations to address the risks associated with use of LMMs in health care.

The responsibilities, detailed below, are defined in government-mandated laws, policies and regulations, as it is governments that must ultimately determine whether an AI-based system should be permitted for use in health care. Developers and providers must also fulfil mutual responsibilities if an LMM is to be used in health care. Such responsibilities could be defined by governments or negotiated between the two parties through a contract if laws have not been written or updated to account for them.

Major risks that must be addressed before deployment include system-wide bias, false information or hallucinations for health-care uses, privacy of data entered into an LMM, manipulation and automation bias.

5.2 Measures that governments can introduce to address such risks and ethical principles that should be upheld

The speed of development of LMMs and of applications that include an LMM requires that governments rapidly develop regulations and specific criteria for using these AI algorithms in health-care systems and for other scientific and medical purposes. The approach should consist of assessment and approval of AI technologies intended for use in health care or medicine by a regulatory agency, such as a medical device or pharmaceutical agency, although governments could establish a new agency for this purpose. One challenge for low- and middle-income countries is that their regulatory agencies are already under-resourced and overwhelmed by pharmaceutical regulation.

The government of at least one high-income country has agreed with the largest technology companies that its foundation models will be assessed in a voluntary public evaluation, with disclosure of outcomes to provide information to the public and researchers about the models and to encourage the companies to correct any errors (146); however, a voluntary approach is likely to be neither sufficient nor sustainable.

Evaluation of LMMs and applications should not address only the AI systems or algorithms used in the health-care system, as there are also significant risks associated with the use of LMMs and applications in a grey zone between clinical and “wellness” applications. Given the rapid proliferation of such technologies, governments should, at least initially, identify such

applications, set common standards and regulations and prohibit applications that do not meet the standards and regulations from being deployed to the public.

Developers and providers should bear the burden of proof, when required, to demonstrate that an AI technology intended for use in health care meets the minimum requirements set out in a law or policy. It should not be assumed that, given the known risks and challenges associated with LMMs, AI algorithms and applications with an LMM are safe and effective or that they are superior to AI or non-AI based approaches that are already in wide use.

Several laws, policies and cross-cutting requirements that could apply to use of LMMs in health care and medicine are described below.

Disclosure (transparency) requirements: Appropriate regulation requires not only that governments have the capacity and discretion to decide what they can assess and approve for use but also adequate information for conducting such an evaluation. Disclosure is necessary to both adequately regulate an AI technology and to ensure that other actors in the AI value chain can use the technology safely. For example, unless a developer discloses the performance of a general-purpose foundation model (such as its propensity to hallucinate), a provider may not have the necessary information to fine-tune the model or to avoid marketing the technology. Such forms of disclosure by a provider or developer can also assist users, such as medical providers, in deciding not to use an LMM that could provide incorrect information or to scrutinize outputs more carefully.

Disclosure and transparency are WHO guiding principles, as well as measures to improve the “explainability” and intelligibility of an AI-based system, and should be required in the assessment of a general-purpose foundation model or application. The WHO guidance on the ethics and governance of AI for health (1) recommended that “government regulators should require the transparency of certain aspects of an AI technology, while accounting for proprietary rights, to improve oversight and assessment of safety and efficacy. This may include an AI technology’s source code, data inputs and analytical approach”. New forms of disclosure that are relevant for LMMs may include their performance in internal testing and their carbon and water footprints. Standards may also be required for “open weights”, which allow regulators, other developers, civil society and providers to understand the outputs of training an algorithm, or the knowledge that an LMM has obtained during its training (147,148).

Several forms of disclosure could assist a provider, user or regulator, including describing the capability and limitations of a foundation model, evaluation of the model according to public or industry standard benchmarks and reporting the results of internal and external testing of the model and its optimization (134). Disclosures, particularly of the risks that may be associated with an LMM or application, could be advertised clearly, and has been compared by one researcher to a “nutrition label” (129).

Data protection laws: The development of LMMs and how developers manage the data required to train an LMM may violate data protection laws. A separate problem is that data entered into an LMM or application to produce a specific output that may include sensitive

personal information may be disclosed either accidentally or through prompts. Potential disclosures are the reason that many large companies, including technology companies that are developing and commercializing LMMs, prohibit their own employees from using such algorithms (149).

Disclosures of data violate the developer's responsibility to protect autonomy. Developers may also violate data protection laws if sensitive data are held longer than permitted according to data minimization requirements (85). One developer allows users to opt out of any content they supply for refining its chatbot's performance (150). Governments that permit use of LMMs should ensure that they set, extend and enforce data protection rules to cover the data entered into an LMM. The Chinese Government's regulation for LMMs includes such a requirement, although the protection applies only to users in China (151).

Assessment of general-purpose foundation models and/or applications used in health care: human rights law versus risk-based frameworks: Several legislative frameworks are being developed to evaluate and regulate AI technologies. One question with respect to such frameworks is whether AI technologies must satisfy human rights obligations ("fundamental rights" according to the European Union) or whether a different approach should be used, to assess AI technologies in a risk-based framework. The European Union, under the AI Act, has adopted a risk-based framework (152). A risk-based framework, it is argued, could help to identify the requirements or burden of proof that must be provided for a technology, the burden of proof increasing with the level of risk of the technology.

All AI systems or tools used in health care and medicine should have to respect ethical obligations and human rights standards that affect, for example, a person's dignity, autonomy or privacy. These include general-purpose foundation models. Human rights and ethical principles are non-negotiable and must be upheld, irrespective of the risk associated with an AI technology or the benefit it may confer (153). The fact that an AI algorithm is considered "low risk" does not exempt it from scrutiny, and a developer or provider should ensure the algorithm respects human rights and ethical obligations. A human rights impact assessment can be conducted to determine whether an LMM or an application adheres to such commitments and can therefore be used safely.

The WHO guidance on the ethics and governance of AI for health (1) recommended that "governments should enact laws and policies that require government agencies and companies to conduct impact assessments of AI technologies, which should address ethics, human rights, safety and data protection, throughout the life cycle of an AI system". The guidance also noted that "impact assessments should be audited by an independent third party before and after introduction of an AI technology and published" (1). The results of impact assessments should be disclosed publicly, while accounting for proprietary or sensitive information, and should be available to the public and groups that may be affected. In the same way as for audits (see above), impact assessments might have to be examined closely, especially if they are conducted by third parties that offer tools or services, to ensure that they are of adequate quality and rigour.

Impact assessments can reveal, for example, whether an AI technology might introduce system-wide bias, risk the privacy of users who share personal data or lead to manipulation of a user. Risks to privacy should be addressed by collaboration between providers and developers in developing LMMs that preserve individual privacy. Work on such an LMM is under way by one company and a hospital system in the USA, although the project is considered unlikely to succeed because the data cannot be fully de-identified (154). Impact assessments can also ensure that use of a general-purpose foundation model or application maintains humans in the loop to avoid automated decision-making in which a user receives false information or misinformation or a health-care provider or patient relies uncritically on the output of an LMM, which would be a form of automation bias.

Governments may instead choose to use a risk-based framework for LMMs to be used in health care and medicine. For those functions that are considered to be higher risk, such as providing a prescription or mental health advice to a person with severe depression or use of an AI technology by a vulnerable or marginalized population, the burden of proof will be higher. There is concern that, if a government selects a risk-based approach, it will be considered sufficient or be used as a substitute for a human rights-based approach (153). A risk-based framework might exclude certain LMMs or applications from evaluation, which may appear to be low risk but could ultimately lead to harm.

Additional questions include whether an AI regulatory assessment should apply to foundation models regardless of their eventual use, whether the assessment should apply only to the largest, most widely used foundation models (“systemic foundation models”⁷) and when such assessments should apply to providers.

This guidance does not include a recommendation on whether all foundation models, regardless of how they are to be used, should be subjected to a risk-based and/or rights-based assessment. The guidance also does not recommend whether an AI regulatory assessment of general-purpose foundation models should apply only to the largest (systemic) LMMs. The expert group did note one concern about assessments designed to apply to all LMMs, irrespective of size, which is that it could “freeze in place” the dominance of the largest companies, as the standards may be such that only those companies can feasibly comply with them or that the standards best suit their business model and aims (155). This concern has gained the attention of competition authorities, which have found that first-movers may use “unfair methods of competition to entrench their current power or use that power to gain control over a new generative AI market” (141). Competition authorities are expected to exercise greater scrutiny of the use of LMMs, although they focus on the practices used by companies that develop LMMs (156).

Providers should also be subject to AI regulatory assessments, as their use of an LMM may change its purpose and function from those determined by the developer but is controlled by the provider. Thus, if a general-purpose foundation model is adapted for use in health care or medicine by a provider, to which the developer agrees, both the developer and the provider

7 Presentation by Kai Zenner, European Parliament, Head of Office for Axel Voss, Conference on AI for Good, 5 June 2023.

should comply with requirements for use of LMMs in health care and medicine. The regulatory burden on providers should be greater if their use of a product or application diverges substantially from or changes the foundation model in ways that are beyond the control of the developer.

Medical device regulation: A government may determine that a general-purpose foundation model or application qualifies as a medical device. While there is little guidance on which LMMs qualify as medical devices, one regulator said that “LMMs only directed toward general purposes and whose developers make no claim that their software can be used for a medical purpose are unlikely to qualify as medical devices” (157). The regulator also noted, however, that: “LMMs that are developed for, or adapted, modified or directed toward specifically medical purposes are likely to qualify as medical devices. Additionally, where a developer makes claims that their LMM can be used for a medical purpose, this again is likely to mean the product qualifies as a medical device” (157).

Chatbots based on LMMs that provide medical advice are likely to be characterized as medical devices under current European Union and US regulatory standards (158). The WHO guidance on the ethics and governance of AI for health (1) recommended that: “government regulators should require that an AI system’s performance be tested, and sound evidence obtained from prospective testing in randomized trials and not merely from comparison of the system to existing datasets in a laboratory”.

If an LMM or application is to be regulated as a medical device, the developer and/or provider should bear the burden of proof by providing evidence that the device performs as marketed and that it meets the requirements of current or amended national laws. This may include various requirements, such as adhering to ethical obligations related to bias and privacy. Newly proposed regulations on AI technologies for medical devices in the European Union and the USA will probably integrate ethical principles related to the use of AI in health, including “explainability”, control of bias and transparency. It is unlikely that current chatbots that include LMMs could meet such standards (158).

LMMs for clinical decision support are already being used experimentally. Although these LMMs include disclaimers, they do not obviate application of medical device laws, “which dictate that such experiments should take place only in an authorized clinical trial setting under appropriate controls to protect patients and to produce clinically relevant outcomes” (158). Governments could examine controlled experimental uses of such LMMs in regulatory “sandboxes”, which would allow testing in a live environment in actual clinical settings with safeguards and oversight to protect health systems from risks or unintended consequences. Such use may, however, be appropriate only in countries in which new health-care products and services and their specifications are subject to formal regulation and data protection regulations (1).

Consumer protection law: Governments should develop and use consumer protection laws to ensure that any negative consequences of LMMs and applications do not reach users and patients. Consumer protection laws could be applied, for example, to prevent practices that

would be tantamount to manipulation (159). In the USA, several government departments and agencies are applying consumer protection laws and other regulations to prevent discrimination and bias in automated systems (159). Such laws can enable governments to require entities that seek to commercialize such technologies to address the causes of any negative consequences and to protect patients and their families from any current or future harm (93). Consumer protection laws, or other regulations, could be used to require that LMMs and applications are restricted in use of language that could misdirect or mislead an end-user into ascribing human-like qualities to an LMM. Such laws could therefore restrict use of or prevent LMMs or applications from using words such as “I think”, “I suppose” or “I suggest”.

Recommendations:

- Governments should, as resources permit, assign an existing or new regulatory agency to assess and approve LMMs and applications intended for use in health care or medicine.
- Certain aspects of an LMM and its applications should be transparent to allow oversight and assessment of its safety and efficacy by regulators. This may include the source code, data inputs, model weights and analytical approach. Additional forms of disclosure to be considered by a government are the performance of an LMM or application in internal testing and its carbon and water footprints.
- Governments should ensure that data protection rules apply to data entered into an LMM or application by a user.
- Government laws, policies and regulations should ensure that LMMs and applications used in health care and medicine, irrespective of the risk or benefit associated with the AI technology, meet ethical obligations and human rights standards that affect, for example, a person’s dignity, autonomy or privacy.
- Governments should enact laws and policies that require providers and developers to conduct impact assessments of LMMs and applications, which should address ethics, human rights, safety and data protection, throughout the life cycle of an AI system. The impact assessments should be audited by an independent third party before and after introduction of an AI technology and should be in the public domain.
- The regulatory burden on a provider should increase if the product or application substantially diverges from or changes the foundation model in ways that are out of the control of the developer of the model.
- Governments should ensure that, for an LMM or application that is regulated as a medical device, the developer and/or provider is responsible for the burden of proof that the device performs as marketed and that it meets the requirements of the country’s laws or amended laws.

- Governments should ensure that LMMs or applications for supporting clinical decisions that are not yet approved for use not be used on an experimental basis outside an authorized clinical trial setting. Governments may facilitate controlled experimental uses of LMMs through regulatory “sandboxes”, which allow testing in a live environment in actual clinical settings, with safeguards and oversight to protect the health system from risks or unintended consequences.
- Governments should use consumer protection laws to ensure that any negative consequences of use of LMMs and applications do not affect users, including patients. Consumer protection laws could be applied, for example, to prevent practices that would be tantamount to manipulation or to address the causes of other negative consequences of LMMs or applications in order to protect patients and their families from any current or future harm.

6 Deployment with general-purpose foundation models (LMMs)

Even when an LMM or an application with an LMM has been ethically designed and undergone appropriate regulatory scrutiny, it may still carry risks when commercialized. The deployer of an AI health-care application or tool could be either the developer or the provider of an LMM or application or, for example, a ministry of health, a hospital, a health-care company or a pharmaceutical company.

6.1 Risks to be addressed when deploying a health-care service or application with a general-purpose foundation model (LMM)

Risks during deployment may be due to the unpredictability of LMMs and the responses they provide, the possibility of use of a general-purpose foundation model in a manner that was not anticipated by either the developer or the provider, and because responses generated by an LMM may change over time.

Major risks that must be addressed when deploying an LMM are:

- inaccurate or false responses,
- bias,
- privacy of data entered into and put out by an LMM,
- accessibility and affordability of an LMM,
- impacts on labour and employment,
- automation bias and skills degradation, and
- the quality of interactions between health-care providers and patients.

This section describes how the actors in the AI value chain, including users, can mitigate or prevent risks and the role of governments in regulating the use of AI tools once an LMM has been deployed, while equipping and training health-care workers and other actors in health systems to maximize appropriate use of an LMM.

6.2 On-going responsibilities of developers and providers during deployment

Developers and providers have responsibilities and obligations even after an LMM or application is approved for use, either because the developer or provider deploys the LMM or because certain risks can be addressed after deployment only by a developer or provider. Such obligations might have to be required by regulations or laws to ensure that developers and providers allocate adequate resources and attention.

First, governments should introduce mandatory post-release auditing and impact assessments, including for data protection and human rights, by independent third parties when an LMM is deployed on a large scale (155,160). Post-release auditing and impact assessments should be published and should address outcomes and impacts disaggregated by type of user, for example by age, race or disability.

Secondly, governments could hold providers or developers responsible for inaccurate, false or toxic content from an LMM after its release that neither the provider nor the developer took steps to correct or avoid. The Chinese Government's regulation on generative AI, for example, stipulates that it must not produce information that is "false and harmful" (151), with possible enforcement by the Government. In the European Union, addition of an LMM to a product or service could create additional responsibilities for the developer and the provider of the LMM. For example, if an LMM is integrated into a service that is within the scope of regulation of digital services, such as the European Union Digital Services Act, the LMM would be indirectly subject to regulatory scrutiny, which could require regulatory oversight because of the tendency of LMMs to hallucinate (120).

Thirdly, a developer and provider might be required to provide ongoing operational disclosures in order for governments and users to use an LMM safely. These could include sufficient technical documentation (133,134).

6.3 Responsibilities of deployers

Deployers are also responsible for avoiding or mitigating risks associated with use of an LMM or application.

First, a deployer should use information from developers or providers to decide not to use an LMM or application in an inappropriate setting, because of biases in the training data, contextual bias that renders the LMM inappropriate for the setting or other avoidable errors or potential risks known to the deployer. If a deployer receives clear, adequate warning of such risks and still offers the LMM for use in inappropriate settings, the deployer should be held accountable for any resulting harm.

Secondly, deployers should communicate any risks that they should reasonably know could result from use of an LMM and any errors or mistakes that have harmed users. Such warnings should not be in fine print or easy to miss. In some circumstances, a deployer may be responsible, even if not required by a law or regulation, for suspending use or removing an LMM or application from the market to avoid harm.

Thirdly, deployers can take steps to improve the affordability and accessibility of an LMM. A deployer can ensure that pricing or subscription fees for use of an LMM correspond to the capacity of a government or other user to pay and should ensure that appropriate LMMs are trained and provided in languages and scripts that can be used by people who are otherwise ignored or excluded from the benefits of technology. Deployers should also request providers and developers to ensure that current and future LMMs are available in several languages.

6.4 Government programmes and practices

Introduction of LMMs into health-care systems and for other health-care-associated uses will require significant adjustment by health-care professionals. Neither developers nor providers have the interest, resources or expertise to ensure appropriate use of an LMM by health-care professionals or for other uses that involve individuals with specialized training and/or expertise.

As in the design of a general-purpose foundation model (see above), governments could enlist both health-care professionals and patients in “human oversight colleges” to ensure that new LMMs and applications used in clinical decision-making are used appropriately and do not undermine the rights of patients.⁸

Governments, universities (health science faculties) or health-care providers such as in hospitals can also ensure that health workers use an LMM to deliver clinical care effectively and are appropriately trained in other uses. Health-care professionals and clinicians should be trained: (i) to understand how LMMs make decisions and the limits of understanding how such decisions are made, (ii) to identify concern about appropriate use, (iii) in methods for avoiding automation bias, (iv) engaging with and educating patients who may be or are considering use of LMMs, and (v) cybersecurity risks associated with the use of LMMs (161).

Training and continuing education of health workers is of particular importance for informing patients, laypeople and other third parties when advice is generated by an LMM or that information provided by an LMM has been used by the provider in making a medical decision or for another medical function. In such notifications, the patient or layperson should be fully apprised of the risks associated with use of LMMs to preserve his or her right to informed consent.

⁸ Communication from David Gruson, WHO expert on ethics and governance of AI for health.

Health-worker training is also critical to ensure that, when they use an LMM professionally, their duties do not unknowingly violate laws, especially those related to the protection of health data and information. For example, medical providers who introduce “protected health information” into an LMM chatbot may be violating laws such as the Health Insurance Portability and Accountability Act in the USA (150). As popular LMMs become “trusted” by health-care workers, for example, they may disclose more patient data than they realize (154).

Other stakeholders in a health-care system should be educated on the benefits, risks, uses and challenges of LMMs in health care and how LMMs differ from other technologies for generating information or advice and how they been used for other purposes in health care. Broader public awareness of the use of LMMs in health care and other domains should be improved. WHO guidance on the ethics and governance of AI for health (1) recommends that: “The public should be engaged in the development of AI for health to understand forms of data sharing and use, to comment on the forms of AI that are socially and culturally acceptable and to fully express their concerns and expectations. Further, the general public’s literacy in AI technology should be improved to enable them to determine which AI technologies are acceptable”.

Governments that supply an LMM or application to a health system could use their procurement authority to foster certain practices among developers, providers and deployers. Procurement of a critical LMM or application for use in a health-care system can eliminate barriers to access and affordability if the AI technology does not displace other health-care investments that may be more effective, equitable and affordable. Public procurement can establish requirements for transparency with respect to data training, quality assurance, risk assessment, mitigation and external audits. Such requirements may be critical if a country has neither relevant legislation nor a regulatory agency with the resources to regulate LMMs effectively.

Recommendations:

- Governments should introduce mandatory post-release auditing and impact assessments, including for data protection and human rights, by independent third parties when an LMM is deployed on a large scale. The auditing and impact assessments should be published and should include outcomes and impacts disaggregated by the type of user, including for example by age, race or disability.
- Governments could hold providers or developers responsible for inaccurate, false or toxic content issued by an LMM after its release, which has not been corrected or avoided by either the provider or the developer.
- Governments should require ongoing operational disclosures by both developers and providers to ensure that LMMs and applications can be used safely. These could include sufficient technical documentation.

- In accordance with information obtained from either developers or providers, deployers should not use an LMM or application in a setting that is inappropriate because of biases in the training data, contextual bias that renders the LMM inappropriate for a particular setting, or other potential errors or risks, such as inaccurate, false or toxic content published by an LMM, that are known to the deployer and can be avoided.
- Deployers should communicate any risks that they should reasonably know could result from use of an LMM, as well as errors that have caused harm to users; such warnings should not be in fine print (or easy to miss). In some circumstances, a deployer may be responsible, even if not required by a law or regulation, to suspend use of or to remove the LMM or application from the market to avoid future harm.
- Deployers should improve the affordability and accessibility of an LMM, by ensuring that pricing or subscription fees for use of an LMM are in line with the capability of a government or other user to pay, and should ensure that appropriate LMMs are trained and offered in languages and scripts that reach people who are otherwise ignored or excluded from the benefits of technology. Deployers should request providers and developers to ensure that current and future LMMs are developed with various languages.
- Governments should facilitate the participation of health-care professionals and patients, in “human oversight colleges” to ensure that new LMMs and applications used in making clinical decision are used appropriately and do not undermine the rights of patients.
- Ministries of health and universities (health science faculties) should train health-care professionals and clinicians: (i) to understand how LMMs make decisions (and the limits of understanding how such decisions are made), (ii) to identify and understand concerns about appropriate use, (iii) in methods to avoid automation bias, (iv) to engage with and educate patients who may be or are considering use of LMMs, and (v) on the cybersecurity risks associated with use of LMMs.
- Governments, health service providers, health researchers and funders should engage the public so that they understand different forms of data-sharing and use, can comment on whether and how LMMs are socially and culturally acceptable and can fully express their concerns and expectations. Further, the literacy of the public in AI technology should be improved to enable them to identify acceptable uses and types of LMMs.
- Governments that supply an LMM or application through a health system should ensure that their procurement authority fosters certain practices by developers, providers and deployers, including transparency.

7 Liability for LMMs

As LMMs gain broader use in health care and medicine, errors, misuse and ultimately harm to individuals are inevitable. Liability rules will have to be used to compensate individuals for such harm, with establishment of new forms of redress when current approaches are insufficient or out of date.

The design, development, quality assurance and deployment of AI technologies involve various entities, each of which plays a distinct role. This can complicate assignment of liability. Developers may demand that downstream entities, such as providers and deployers, be liable for any harm resulting from use of an LMM, while downstream entities may claim that previous actions, such as choice of the data used to train an algorithm, are the cause. Developers and providers may also claim that, once a medical AI technology has been approved for use by a regulator, they should no longer be held liable for harm (regulatory pre-emption) (1). Establishment of liability along the value chain is a challenge for lawmakers and policy-makers.

A critical function of rules for civil liability is to ensure that a victim of damage can claim compensation and redress, no matter how difficult it may be to assign blame and responsibility among the entities involved in development and deployment of an AI technology. If victims find it too difficult to obtain compensation, there can be no justice and no incentive for parties in the AI value chain to avoid such harm in the future. The rules should also ensure that the compensation is adequate for the harm suffered.

The European Union in its proposed AI Liability Directive simplifies the burden of proof by a victim by introducing a “presumption of causality” (162). Thus, if a victim can demonstrate that one or more entities did not comply with an obligation relevant to the harm and that a causal link with AI performance is likely, the court can presume that non-compliance was the cause of the damage (162). The onus is thus placed on the liable party to rebut the presumption, for example by indicating another party as the cause of the damage. The scope of the legislation is not limited to the original maker of an AI system but includes any participant in the AI value chain (162). When all the actors in the AI value chain are held jointly liable, they can demonstrate their effectiveness in assessing and mitigating risks in order to reduce their liability.

It is still possible, however, that a liability regime does not provide full clarity and redress for injuries caused by AI-driven products and services, especially if an individual does not know that a LMM was used in making a medical decision. New rules may leave gaps in liability for injuries caused by AI-driven medical technologies (163). As LMMs are highly speculative, poorly understood and being rushed to market, governments may wish to consider LMMs used in health care as products for which developers, providers and deployers will be held to a strict liability standard. Holding these actors accountable for any error might ensure that a

patient will be compensated if the error affects them (1), although this depends on whether the patient knew that an LMM was used. While such continuing liability might discourage the use of increasingly sophisticated LMMs, it might also temper a willingness to take unnecessary risks and to deploy new LMMs into health care or public health settings before their many risks and potentials harms have been fully identified and addressed (1).

A liability regime for AI might not, however, be adequate to assign fault, as algorithms are evolving in ways that neither developers, providers nor deployers can fully control. Furthermore, there may be situations or jurisdictions in which a person who is harmed is unable to recover damages. For example, in the USA, a patient who is injured when using an LMM directly to seek advice may not be able to recover damages because AI systems themselves are not included in professional liability rules, and exceptions or limitations to product or consumer liability laws may preclude recovery (163). In other areas of health care, compensation is occasionally provided without assignment of fault or liability, such as for medical injuries resulting from adverse effects of vaccines. The original WHO guidance recommended determination of “whether no-fault, no-liability compensation funds are an appropriate mechanism for providing payments to individuals who suffer medical injuries due to the use of AI technologies, including how to mobilize resources to pay any claims” (1). That recommendation is also valid today and could be a means for determining compensation for injuries caused by LMMs or applications with LMMs.

Recommendation:

- Governments should establish liability along the value chain of the development, provision and deployment of LMMs and applications to ensure that a victim of damage can claim compensation, irrespective of the difficulty of assigning blame and of the responsibilities of the different entities involved in the development and deployment of the technology.

8 International governance of LMMs

Governments should support collective development of international rules for the governance of LMMs and other forms of AI used in health care, as such uses are proliferating globally. One example is the WHO global strategy on digital health 2020-2025. The process should include greater cooperation and collaboration within the United Nations system to respond to the opportunities and challenges of deploying AI in health care and of its wider application in society and the economy. Unless governments work together to set appropriate, enforceable standards, the number of LMMs and other forms of AI that do not meet appropriate legal, ethical and safety standards will increase, potentially causing harm if regulations and other types of protection are not introduced or are not properly enforced, whether willingly or because there are inadequate resources. WHO recently issued a new publication, in consultation with regulatory agencies worldwide, that outlines key principles that governments and regulatory authorities can follow to develop new guidance or adapt existing guidance on AI (164).

International governance can avoid a “race to the bottom” among companies seeking a first-mover advantage in which standards of safety and efficacy are ignored, and among governments seeking advantage in the geopolitical race for technological supremacy. Thus, international governance can ensure that all companies meet minimum standards of safety and efficacy and also avoid introduction of regulations that provide a competitive advantage or disadvantage for either companies or governments. International governance can make governments accountable for their investments and participation in the development and deployment of AI-based systems and ensure that they introduce appropriate regulations that respect ethical principles, human rights and international law. The absence of globally enforceable standards may also have a negative impact on product adoption.

International governance could take several forms. One suggestion is to establish a public research agency, funded by several governments, such as the European Organization for Nuclear Research, an international collaboration, with funding and human resources to pursue large, transformative projects the results of which are shared openly (165,166). In a separate proposal, it was suggested that such an entity could be charged with developing the most advanced, and most risky, forms of AI in a highly secure facility, making other attempts to build such forms of AI illegal (167). At present, such large-scale projects are not in the domain of publicly funded projects to generate public goods but are the purview of large technology companies in commercial competition with one another. Other leaders, including world leaders and technology executives, have called for AI to be treated similarly to nuclear weapons, with a global regulatory framework similar to treaties for the use of nuclear arms (109).

Whatever the form of international governance that is taken forward, it is imperative that it not be shaped solely by high-income countries or by high-income countries that work mostly or solely with the world's largest technology companies (168). Standards developed by and for high-income countries and technology firms, whether for all applications of AI or for specific use of LMMs in health care and medicine, will leave most of humanity, in low- and middle-income countries, with no role or voice in shaping the standards. This would render future AI technologies potentially dangerous or ineffective in the very countries that might ultimately benefit the most.

International governance of AI could require that all stakeholders cooperate through networked multilateralism, as proposed by the United Nations Secretary-General in 2019 (169), which would bring together the United Nations family, international financial institutions, regional organizations, trading blocs and others, including civil society, cities, businesses, local authorities and young people, to work more closely, effectively and inclusively. Placing ethics and human rights at the centre of the development and deployment of LMMs could make a substantial contribution towards achievement of universal health coverage.

Recommendation:

- Governments should support collective development of international rules for the governance of AI. Whatever the form of governance, it must not be shaped solely by high-income countries or by high-income countries working mostly or solely with the world's largest technology companies, as that approach would leave most of humanity, in low- and middle-income countries, without a role or voice in shaping international governance of AI.

References

1. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021 (<https://www.who.int/publications/i/item/9789240029200>, accessed 26 May 2023).
2. Khullar D. Can A.I. treat mental illness? *The New Yorker*, 27 February 2023 (<https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness>, accessed 29 May 2023).
3. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773–84. doi:10.1038/s41591-022-01981-2.
4. Hariri Y, Harris T, Raskin A. You can have the blue pill or the red pill, and we're out of blue pills. *The New York Times*, 24 March 2023 (<https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html>, accessed 26 May 2023).
5. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ et al. Foundation models for generalist medical artificial intelligence, *Nature*. 2023;616(7956):259–65. doi:10.1038/s41586-023-05881-4.
6. Hu K. Chat GPT sets record for fastest growing user-base – analyst note. *Reuters*, 2 February 2023 (<https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>, accessed 26 May 2023).
7. Weise K, Grant N. Microsoft and Google unveil A.I. tools for businesses. *The New York Times*, 16 March 2023 (<https://www.nytimes.com/2023/03/16/technology/microsoft-google-ai-tools-businesses.html>, accessed 26 May 2023).
8. Yang Z. Chinese tech giant Baidu just released its answer to ChatGPT. *MIT Technology Review*, 16 March 2023 (<https://www.technologyreview.com/2023/03/16/1069919/baidu-ernie-bot-chatgpt-launch/>, accessed 26 May 2023).
9. Murgia M, Bradshaw T. Musk to launch AI start-up to rival ChatGPT. *Financial Times*, 15 April 2023 (<https://www.ft.com/content/2a96995b-c799-4281-8b60-b235e84aefe4>, accessed 26 May 2023).
10. Heaven WD. The open-source AI boom is built on Big Tech's handouts. How long will it last? *MIT Technology Review*, 12 May 2023 (<https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/>, accessed 26 May 2023).

11. Martin A. Google CEO Sunder Pichai admits people don't fully understand how chatbot AI works, Evening Standard, 17 April 2023 (<https://www.standard.co.uk/tech/google-ceo-sundar-pichai-understand-ai-chatbot-bard-b1074589.html>, accessed 26 May 2023).
12. Roose K. A conversation with Bing's chatbot left me deeply unsettled. The New York Times, 16 February 2023 (<https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html>, accessed 26 May 2023).
13. Marcus G. AI platforms like ChatGPT are easy to use but potentially dangerous, Scientific American, 19 December 2022 (<https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/>, accessed 26 May 2023).
14. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E et al. Sparks of artificial general intelligence: early experiments with GPT-4. ArXiv:2302.12712.
15. McGowran L. OpenAI criticised for lack of transparency around ChatGPT-4. Silicon Republic, 16 March 2023 (<https://www.siliconrepublic.com/machines/openai-gpt4-transparency-ai-concerns-stripe-chatgpt>, accessed 26 May 2023).
16. Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis)informs us better than humans. Sci Adv. 2023;9(26):eadh1850. doi:10.1126/sciadv.adh1850.
17. Volpicelli G. ChatGPT broke the EU plan to regulate AI Politico, 3 March 2023 (<https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/>, accessed 26 May 2023).
18. Arcesati R, Chang W. China is blazing a trail in regulating Generative AI – on the CCP's terms. The Diplomat, 28 April 2023 (<https://thediplomat.com/2023/04/china-is-blazing-a-trail-in-regulating-generative-ai-on-the-ccps-terms/>, accessed 26 May 2023).
19. Martindale J. These are the countries where ChatGPT is currently banned. Digital Trends, 12 April 2023 (<https://www.digitaltrends.com/computing/these-countries-chatgpt-banned/>, accessed 26 May 2023).
20. Johnson K. ChatGPT can help doctors – and hurt patients. Wired, 24 April 2023 (<https://www.wired.com/story/chatgpt-can-help-doctors-and-hurt-patients/>, accessed 28 May 2023).
21. Topol E. Multimodal AI for medicine, simplified. Ground Truths, 14 March 2023 (<https://erictopol.substack.com/p/multimodal-ai-for-medicine-simplified>, accessed 28 May 2023).
22. Heaven WD. AI hype is built on high test scores. Those tests are flawed. MIT Technology Review, 30 August 2023 (<https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/>, accessed 1 October 2023).

23. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–80. doi:10.1038/s41586-023-06291-2.
24. Kulkarni PA, Singh H. Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. *JAMA*. 2023;330(4):317–8. doi:10.1001/jama.2023.11440.
25. Subbamaran N. ChatGPT will see you now: Doctors using AI to answer patient questions. *Wall Street Journal*, 28 April 2023 (<https://www.wsj.com/articles/dr-chatgpt-physicians-are-sending-patients-advice-using-ai-945cf60b>, accessed 28 May 2023).
26. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023; 183(6):589–96. doi: 10.1001/jamainternmed.2023.1838.
27. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New Engl J Med*. 2023;388(13):1233–9. doi: 10.1056/NEJMSr2214184.
28. The potential of large language models in healthcare: Improving quality of care and patient outcomes, *Medium*, 7 December 2022. (<https://medium.com/@BuildGP/the-potential-of-large-language-models-in-healthcare-improving-quality-of-care-and-patient-6e8b6262d5ca>, accessed 28 May 2023).
29. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. 2017;1711.05225v3. doi:10.48550/arXiv.1711.05225.
30. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C et al. A large language model for electronic health records. *npj Digit Med*. 2022;5:194. doi:10.1038/s41746-022-00742-2.
31. Ghahramani Z. Introducing PaLM 2. The Keyword, 10 May 2023 (<https://blog.google/technology/ai/google-palm-2-ai-large-language-model/>, accessed 28 May 2023).
32. Weise K, Metz C. When A.I. chatbots hallucinate. *The New York Times*, 9 May 2023 (<https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>, accessed 1 June 2023).
33. Bender EM, Gebru T, McMillan-Major A, Mitchell M. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 2021, pp. 610–23. doi: 10.1145/3442188.3445922.
34. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333–5. doi:10.1016/s2589-7500(23)00083-3.

35. Metz, Cade. Chatbots may 'hallucinate' more often than many realize, *The New York Times*, 6 November 2023 (<https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html>, accessed 7 November 2023).
36. Acar OA. AI prompt engineering isn't the future, *Harvard Business Review*, 6 June 2023 (<https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future?registration=success>, accessed 26 June 2023).
37. GPT-4 system card. Open AI, 23 March 2023 (<https://cdn.openai.com/papers/gpt-4-system-card.pdf>, accessed 28 May 2023).
38. GPT-4. OpenAI, 14 March 2023 (<https://openai.com/research/gpt-4>, accessed 28 May 2023).
39. Radford A, Kleinman Z. ChatGPT can now access up-to-date information. *BBC News*, 27 September 2022 (<https://www.bbc.com/news/technology-66940771>, accessed 1 October 2023).
40. Kruge S, Ostermaier A, Uhl M. The moral authority of ChatGPT. *ArXiv:23101.07098*. doi:10.48550/arXiv.23101.07098.
41. Mickle T, Metz C, Grant N. The chatbots are here, and the internet industry is in a tizzy, *The New York Times*, 8 March 2023 (<https://www.nytimes.com/2023/03/08/technology/chatbots-disrupt-internet-industry.html>, accessed 29 May 2023).
42. Woo M. Trial by artificial intelligence. *Nature*. 2019;573:S100–2 (<https://media.nature.com/original/magazine-assets/d41586-019-02871-3/d41586-019-02871-3.pdf>, accessed 29 May 2023).
43. Muralidharan V, Burgart A, Daneshjou D, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *npj Dig Med*. 2023;6:166. doi:10.1038/s41746-023-00898-5.
44. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning Individual Differences*. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.
45. Reddy CD, Lopez L, Ouyang D, Zou JY, He B. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *J Am Soc Echocardiogr*. 2023;36(5):482–9. doi:10.1016/j.echo.2023.01.015.
46. Knight W. These ChatGPT rivals are designed to play with your emotions. *Wired*, 4 May 2023 (<https://www.wired.com/story/fast-forward-chatgpt-rivals-emotions/#:~:text=12%3A00%20PM-,These%20ChatGPT%20Rivals%20Are%20Designed%20to%20Play%20With%20Your%20Emotions,%2C%20-companionship%E2%80%94and%20even%20romance.>, accessed 29 May 2023).

47. Smuha NA, De Ketalaere M, Coeckelbergh M, Dewitte P, Pouillet Y. Open letter: We are not ready for manipulative AI – urgent need for action. KU Leuven, 31 March 2023 (<https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai>, accessed 29 May 2023).
48. Cuthbertson A. “No, I’m not a robot”: ChatGPT successor tricks worker into thinking it is human. Independent, 15 March 2023 (<https://www.independent.co.uk/tech/chatgpt-gpt4-ai-openai-b2301523.html>, accessed 26 June 2023).
49. Walker L. Belgian man dies by suicide following exchanges with chatbot, The Brussels Times, 28 March 2023 (<https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>, accessed 29 May 2023).
50. DeGuerin M. Oops: Samsung employees leaked confidential data to ChatGPT. Gizmodo, 6 April 2023 (<https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376>, accessed 29 May 2023).
51. Privacy policy. OpenAI, 27 April 2023 (<https://openai.com/policies/privacy-policy>, accessed 29 March 2023).
52. Coles C. 11% of data employees paste into ChatGPT is confidential. Cyberhaven, 19 April 2023 (<https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/>, accessed 29 May 2023).
53. Mihalcik C. ChatGPT bug exposed some subscribers’ payment info. CNET, 24 March 2023. (<https://www.cnet.com/tech/services-and-software/chatgpt-bug-exposed-some-subscribers-payment-info/>, accessed 29 May 2023).
54. Moodley K, Rennie S. ChatGPT has many uses. Experts explore what this means for healthcare and medical research. The Conversation, 22 February 2023 (<https://theconversation.com/chatgpt-has-many-uses-experts-explore-what-this-means-for-healthcare-and-medical-research-200283>, accessed 2 June 2023).
55. De Proost M, Pozzi G. Conversational artificial intelligence and the potential for epistemic injustice. *Am J Bioethics*. 2023;23(5):51–3. doi:10.1080/15265161.2023.2191020.
56. Disability and employment. New York: United Nations, Department of Economic and Social Affairs (Disability); undates (<https://www.un.org/development/desa/disabilities/resources/factsheet-on-persons-with-disabilities/disability-and-employment.html>, accessed 11 September 2023).
57. Whittaker M, Alper M, Bennett CL, Hendren S, Kaziunas L, Mills Met al. Disability, bias and AI. New York: AI Now Institute; 2019 (<https://ainowinstitute.org/wp-content/uploads/2023/04/disabilitybiasai-2019.pdf>, accessed 11 September 2023).

58. Hallman J. AI language models show bias against people with disabilities, study finds. University Park (PA): Penn State University; 2022 (<https://www.psu.edu/news/information-sciences-and-technology/story/ai-language-models-show-bias-against-people-disabilities/>, accessed 11 September 2023).
59. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*. 2023;90:194512. doi:10.1016/j.ebiom.2023.104512.
60. Lohr S. AI may someday work medical miracles. For now, it helps do paperwork. *The New York Times*, 26 June 2023 (<https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html>, accessed 10 July 2023).
61. Eddy N. Epic, Microsoft partner to use generative AI for better EHRs. *Healthcare IT News*, 18 April 2023 (<https://www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs>, accessed 31 May 2023).
62. Nuance and Microsoft announce the first fully AI-automated clinical documentation application for healthcare. Burlington (MA):Nuance; 2023 (<https://news.nuance.com/2023-03-20-Nuance-and-Microsoft-Announce-the-First-Fully-AI-Automated-Clinical-Documentation-Application-for-Healthcare>, accessed 31 May 2023).
63. Ahn S. The impending impacts of large language models on medical education. *Korean J Med Educ*. 2023;35(1):103–7. doi:10.3946/kjme.2023.253.
64. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefi Bioinformatics*. 2022; 23(6):bbac409. doi:10.1093/bib/bbac409.
65. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. *Drug Discov Today*. 2021;26(1):80–93. doi:10.1016/j.drudis.2020.10.010.
66. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*. 2023;613:612. doi:10.1038/d41586-023-00191-1.
67. Zielinski C, Winker MA, Aggarwal R, Ferris LA, Heinemann M, Lapeña JF Jr et al. Chatbots, generative AI, and scholarly manuscripts. *Overijssel: World Association of Medical Editors*; 2023. (<https://wame.org/page3.php?id=106>, accessed 26 June 2023).
68. Gibbs W. Lost science in the Third World. *Sci Am*. 1995;273(2):92–9 doi:10.1038/scientificamerican0895-92.
69. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. *Nat Rev Phys*. 2023;5:277–80. doi:10.1038/s42254-023-00581-4.

70. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. Geneva: World Health Organization; 2010 (<https://apps.who.int/iris/bitstream/handle/10665/258734/9789241564052-eng.pdf>, accessed 26 June 2023).
71. Morozov E. The true threat of artificial intelligence. The New York Times, 30 June 2023 (<https://www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html>, accessed 2 July 2023).
72. Introducing ChatGPTPlus. San Francisco (CA): Open AI; 2023 (<https://openai.com/blog/chatgpt-plus>, accessed 1 June 2023).
73. The hidden workforce that helped filter violence and abuse out of ChatGPT. Wall Street Journal, 11 July 2023 (<https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413>, accessed 13 July 2023).
74. Firth N. Language models may be able to self-correct biases – if you ask them. MIT Technology Review, 20 March 2023 (<https://www.technologyreview.com/2023/03/20/1070067/language-models-may-be-able-to-self-correct-biases-if-you-ask-them-to/>, accessed 1 June 2023).
75. Khan L. We must regulate A.I. Here’s how. The New York Times, 3 May 2023 (<https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html>, accessed 2 June 2023).
76. Hatzius J, Briggs J, Kodnani D, Pierdomenico G. The potentially large effects of artificial intelligence on economic growth (Briggs/Kodnani). Goldman Sachs Economics Research, 26 May 2023 (https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf, accessed 1 June 2023).
77. Milmo D. AI revolution puts skilled jobs at highest risk, says OECD. The Guardian, 11 July 2023 (<https://www.theguardian.com/technology/2023/jul/11/ai-revolution-puts-skilled-jobs-at-highest-risk-oecd-says>, accessed 12 July 2023).
78. Health and care workforce in Europe: time to act. Geneva: World Health Organization; 2022 (<https://iris.who.int/handle/10665/362379>, accessed 1 June 2023).
79. Health workforce. Geneva: World Health Organization; 2023 (https://www.who.int/health-topics/health-workforce#tab=tab_1, accessed 1 June 2023).
80. Hurst L. OpenAI says 80% of workers could see their jobs impacted by AI. These are the jobs most impacted, Euronews.next, 30 March 2023. (<https://www.euronews.com/next/2023/03/23/openai-says-80-of-workers-could-see-their-jobs-impacted-by-ai-these-are-the-jobs-most-affe>, accessed 1 June 2023).

81. A new era of generative AI for everyone. Dublin: Accenture; 2023 (<https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf>, accessed 1 June 2023).
82. Burgess M. The security hole at the heart of ChatGPT and Bing. *Wired*, 25 May 2023 (<https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security>, accessed 1 June 2023).
83. Heikkila M. Open AI's hunger for data is coming back to bite it. *MIT Technology Review*, 19 April 2023 (<https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/>, accessed 1 June 2023).
84. General Data Protection Regulation, Regulation 2016//679 of the European Parliament and of the Council, 27 April 2016. Strasbourg: European Parliament; 2016 (<https://eur-lex.europa.eu/eli/reg/2016/679/oj>, accessed 27 September 2023).
85. The impact of the General Data Protection Regulation on artificial intelligence (STOA Options Brief). Strasbourg: European Parliament; 2020 ([https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU\(2020\)641530\(ANN1\)_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530(ANN1)_EN.pdf), accessed 26 June 2023).
86. OPC launches investigation into ChatGPT. Ottawa: Office of the Privacy Commissioner of Canada; 4 April 2023 (https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230404/, accessed 1 June 2023).
87. Lomas N. Italy orders Chat GPT blocked citing data protection concerns. *Tech Crunch*, 31 March 2023 (<https://techcrunch.com/2023/03/31/chatgpt-blocked-italy/>, accessed 1 June 2023).
88. ChatGPT: Italian SA to lift temporary limitation if OpenAI implements measures. Rome: Italian Data Protection Authority; 2023 (<https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751#english>, accessed 1 June 2023).
89. Weatherbed J. OpenAI's regulatory troubles are only just beginning. *The Verge*, 5 May 2023 (<https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy>, accessed 1 June 2023).
90. Wiggers K. Open AI's new tool attempts to explain language models' behaviours. *Tech Crunch*, 9 May 2023 (<https://techcrunch.com/2023/05/09/openais-new-tool-attempts-to-explain-language-models-behaviors/>, accessed 1 June 2023).
91. Libeau D. ChatGPT will probably never comply with GDPR. 10 April 2023. (<https://blog.davidlibeau.fr/chatgpt-will-probably-never-comply-with-gdpr/>, accessed 1 June 2023).

92. Lomas N. ChatGPT maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher. TechCrunch, 30 August 2023. (https://consent.yahoo.com/v2/collectConsent?sessionId=3_cc-session_6bdecae4-d7b6-448f-8e26-e7805c03b964, accessed 11 September 2023).
93. Fung B. The FTC should investigate Open AI and block GPT over “deceptive” behaviour, AI policy group claims. CNN, 30 March 2023. (<https://edition.cnn.com/2023/03/30/tech/ftc-openai-gpt-ai-think-tank/index.html>, accessed 2 June 2023).
94. Waters R, Murgia M, Espinoza J. Open AI warns over split with Europe as AI regulation advances. Financial Times, 25 May 2023 (<https://www.ft.com/content/5814b408-8111-49a9-8885-8a8434022352>, accessed 1 June 2023).
95. Technology-facilitated gender-based violence: Making all spaces safe. New York: United Nations Population Fund; 2021 (<https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe>, accessed 1 October 2023).
96. Murgia M. DeepMind reinvents itself for AI counterattack. Financial Times, 2 May 2023 (<https://ft.pressreader.com/v99c/20230502/281724093873699>, accessed 2 June 2023).
97. Schaake M. Regulating AI will put companies and governments at loggerheads, Financial Times, 2 May 2023 (<https://www.ft.com/content/7ef4811d-79bb-4b4f-b28f-b46430f0c9ff>, accessed 2 June 2023).
98. Metz, Cade. Tech giants are paying huge salaries for scarce A.I. talent, The New York Times, 22 October 2017 (<https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html>).
99. Leswing K. Google reveals its newest AI supercomputer, says it beats Nvidia. CNBC, 5 April 2023. (<https://www.cnbc.com/2023/04/05/google-reveals-its-newest-ai-supercomputer-claims-it-beats-nvidia-.html>, accessed 2 June 2023)
100. Ahuja K. Antitrust has role in policing AI landscape. Financial Times, 10 April 2023 (<https://www.ft.com/content/953817f5-5bc4-49e1-b583-977cc4780eca>, accessed 2 June 2023).
101. Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. *Science*. 2023;379(6635):884–6. doi:10.1126/science.ade2420.
102. Røttingen JA, Regmi S, Eide M, Young AJ, Viergever RF, Ardal C et al. Mapping of available health research and development data: What’s there, what’s missing, and what role is there for a global observatory? *Lancet*. 2013;382(9900):1286–307. doi:10.1016/S0140-6736(13)61046-6.

103. A new partnership to promote responsible AI. Google Blogs, 26 July 2023 (<https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/#:~:text=Anthropic%2C%20Google%2C%20Microsoft%20and%20OpenAI%20are%20launching%20the%20Frontier%20Model,development%20of%20frontier%20AI%20models>, accessed 29 July 2023).
104. Fact sheet: Biden–Harris Administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. Washington DC: The White House, 21 July 2023 (<https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>, accessed 29 July 2023).
105. Volpicelli G. Europe pitches AI pact to curtail the booming tech’s risk. Politico, 26 May 2023 (<https://www.politico.eu/article/big-tech-rumble-europe-global-artificial-intelligence-debate-ai-pact/>, accessed 29 July 2023).
106. Grant N, Weise K. In AI race, Microsoft and Google choose speed over caution. The New York Times, 7 April 2023 (<https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html>, accessed 2 June 2023).
107. Center for Research on Foundation Models. The Foundation Model Transparency Index, 2023. (<https://crfm.stanford.edu/fmti/>, accessed 21 October 2023).
108. Schiffer Z, Newton C. Microsoft lays off team that taught employees how to make AI tools responsibly. The Verge, 14 March 2023. (<https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs>, accessed 2 June 2023).
109. Milmo D. Google chief warns AI could be harmful if deployed wrongly. The Guardian, 17 April 2023 (<https://www.theguardian.com/technology/2023/apr/17/google-chief-ai-harmful-sundar-pichai>, accessed 2 June 2023).
110. Fiesler C. AI has social consequences, but who pays the price? Tech companies’ problem with ethical debt. The Conversation, 19 April 2023 (<https://theconversation.com/ai-has-social-consequences-but-who-pays-the-price-tech-companies-problem-with-ethical-debt-203375>, accessed 2 June 2023).
111. Criddle, Cristina and Murphy, Hannah, Meta disbands protein-folding team in shift towards commercial AI, Financial Times, 7 August 2023. (https://www.ft.com/content/919c05d2-b894-4812-aa1a-dd2ab6de794a?accessToken=zwAGBZu-oVVwkdORnAXSuJRIEtOqGt0qtt55Sg.MEQCIA1QQ1iG8KPAAnuDAuPvt-Ngds3OzxL1lt-0FnaVbAQFtAiAZvHnmKD_fABj8ZzLTNXRp1v7V38nTcUf_pPxAPdx16A&sharetype=gift&toKen=3ac5a132-e08e-412e-bc3c-08edea8a7417, accessed 18 September 2023).
112. Ananthaswamy A. In AI, is bigger always better? Nature, 8 March 2023 (<https://www.nature.com/articles/d41586-023-00641-w>, accessed 2 June 2023).

113. Li P. Making AI less “thirsty”: Uncovering and addressing the secret water footprint of AI models. ArXiv. 2023;2304.03271v. doi:10.48550/arXiv.2304.03271.
114. Syed N. The secret water footprint of AI technology. The Markup, 15 April 2023 (<https://themarkup.org/hello-world/2023/04/15/the-secret-water-footprint-of-ai-technology>, accessed 2 June 2023).
115. Livingstone G. It’s pillage: Thirsty Uruguayans blast Google’s plan to exploit water supply. The Guardian, 11 July 2023 (<https://www.theguardian.com/world/2023/jul/11/uruguay-drought-water-google-data-center>, accessed 12 July 2023).
116. Thornhill J. The sceptical case on generative AI. Financial Times, 17 August 2023. (<https://www.ft.com/content/ed323f48-fe86-4d22-8151-eed15581c337>, accessed 11 September 2023).
117. Marcus G. The imminent enshittification of the Internet. Substack, 16 August 2023 (<https://garymarcus.substack.com/p/the-imminent-enshittification-of>, accessed 11 September 2023).
118. Pause giant AI experiments: An open letter. Narberth (PA): Future of Life Institute; 2023 (<https://futureoflife.org/open-letter/pause-giant-ai-experiments/>, accessed 13 June 2023).
119. Perrigo B. DeepMind’s CEO helped take AI mainstream. Now he’s urging caution. Time, 12 January 2023 (<https://time.com/6246119/demis-hassabis-deepmind-interview/>, accessed 13 June 2023).
120. Lomas N. Unpacking the rules shaping generative AI. Tech Crunch, 13 April 2023 (<https://techcrunch.com/2023/04/13/generative-ai-gdpr-enforcement/>, accessed 13 June 2023).
121. Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach Soc Sci Res Netw. 2023. doi:10.2139/ssrn.4361607.
122. Lomas N. Report details how Big Tech is leaning on EU not to regulate general purpose AIs. Tech Crunch, 23 February 2023 (<https://techcrunch.com/2023/02/23/eu-ai-act-lobbying-report/>, accessed 20 June 2023).
123. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. “Everyone wants to do model work, not the data work.”: Data cascades in high-stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021. doi:10.1145/3411764.3445518.
124. Browne G. AI is steeped in Big Tech’s digital colonialism. Wired, 25 May 2023 (<https://www.wired.co.uk/article/abeba-birhane-ai-datasets>, accessed 17 June 2023).

125. Baxter K, Schelsinger, N. Managing the risks of generative AI. *Harvard Business Review*, 6 June 2023 (<https://hbr.org/2023/06/managing-the-risks-of-generative-ai>, accessed 17 June 2023).
126. Samuelson P. Generative AI meets copyright. *Science*. 2023;381(6654):158–61. doi:10.1126/science.adi0656.
127. El-Mhamdi E, Farhadkhani S, Guerraoui R, Gupta N, Hoang L, Pinot R et al. On the impossible safety of large AI models. *arXiv*. 2209.15259v2. doi:10.48550/arXiv.2209.15259.
128. Open AI. GPT-4 technical report. *arXiv*:2303.08774v3. doi:10.48550/arXiv.2302.08774.
129. Murgia M. Open AI’s red team: experts hired to “break” ChatGPT. *Financial Times*, 14 April 2023 (<https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8>, accessed 10 July 2023).
130. Clegg N. Openness on AI is the way forward for tech. *Financial Times*, 11 July 2023 (<https://www.ft.com/content/ac3b585a-ce50-43d1-b71d-14dfe6dce999>, accessed 11 July 2023).
131. Huang S, Toner H, Haluza Z, Creemers R, Webster G. Measures for the management of generative artificial intelligence services (draft for comment) (translation). DigiChina. Palo Alto (CA): Stanford University, Program on Geopolitics; 2023 (<https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/>, accessed 17 June 2023).
132. Ye J. China says generative AI rules to apply only to products for the public. *Reuters*, 13 July 2023 (<https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/>, accessed 13 July 2023).
133. Bommasani R, Klyman K, Zhang D, Liang P. Do foundation model providers comply with the draft EU AI Act? Palo Alto (CA): Stanford University, Human-centered Artificial Intelligence; 2021 (<https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>, accessed 17 June 2023).
134. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Strasbourg: European Parliament; 2023 (https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, accessed 10 July 2023).
135. Beyond ChatGPT: How can Europe become a leader in generative AI? Kaiserslautern: German Research Centre for Artificial Intelligence; 2023. (<https://www.dfki.de/en/web/news/jenseits-von-chatgpt-wie-kann-europa-bei-der-generativen-ki-eine-fuehrungsposition-uebernehmen>, accessed 17 June 2023).

136. Spirling A. Why open-source generative AI models are an ethical way forward for science. *Nature*. 2023;616(7957):413. doi:10.1038/d41586-023-01295-4.
137. Vincent J. Meta’s powerful AI language model has leaked online – What happens now? *The Verge*, 8 March 2023 (<https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse>, accessed 29 July 2023).
138. Maffuli S. Meta’s Llama 2 license is not open source. Open Source Initiative, 20 July 2023 (<https://blog.opensource.org/metals-llama-2-license-is-not-open-source/>, accessed 29 July 2023).
139. Marble A. Software licenses masquerading as open source. *marble.onl*, 1 June 2023 (<http://marble.onl/posts/software-licenses-masquerading-as-open-source.html>, accessed 29 July 2023).
140. Keary T. Report finds 82% of open-source software components “inherently risky”. *Venture Beat*, 17 April 2023 (<https://venturebeat.com/security/report-finds-82-of-open-source-software-components-inherently-risky/>, accessed 8 July 2023).
141. Generative AI raises competition concerns. *Technology blog*, 29 June 2023. Washington DC: Federal Trade Commission; 2023 (<https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns>, accessed 29 July 2023).
142. Wishart-Smith H. Generative AI: cybersecurity friend and foe, *Forbes*, 6 June 2023 (<https://www.forbes.com/sites/heatherwishartsmith/2023/06/06/generative-ai-cybersecurity-friend-and-foe/?sh=4407e0884bd2>, accessed 29 July 2023).
143. Metz C. Researchers poke holes in safety controls of ChatGPT and other chatbots. *The New York Times*, 27 July 2023 (<https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html>, accessed 11 September 2023).
144. Harris T, Freuh S. The complexity of technology’s consequences is going up exponentially, but our wisdom and awareness are not. *Issues in Science and Technology*, 16 May 2023 (<https://issues.org/tristan-harris-humane-technology-misinformation-ai-democracy/>, accessed 19 June 2023).
145. Schyns C. The lobbying ghost in the machine: Big Tech’s covert defanging of Europe’s AI Act. Brussels: Corporate Europe Observatory; 2023 (<https://corporateeurope.org/en/2023/02/lobbying-ghost-machine>, accessed 17 June 2023).
146. Fact sheet: Biden–Harris Administration announces new actions to promote responsible AI innovation that protects Americans’ rights and safety. Washington DC: White House, 4 May 2023 (<https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/>, accessed 19 June 2023).

147. Sijbrandij Sid. AI weights are not “open source”. Open Core Ventures, 27 June 2023. (<https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/>, accessed 29 July 2023).
148. Meeker H. Towards an open weights definition. Copyleft Currents, 8 June 2023 (<https://heathermeeker.com/2023/06/08/toward-an-open-weights-definition/>, accessed 29 July 2023).
149. Dastin J, Tong A. Google, one of AI’s biggest backers, warns its own staff about chatbots. Reuters, 15 June 2023 (<https://www.reuters.com/technology/google-one-ais-biggest-backers-warns-own-staff-about-chatbots-2023-06-15/>, accessed 9 July 2023).
150. Kanter GP, Packel EA. Health care privacy risks of AI chatbots. *JAMA*. 2023;330(4):311–2. doi:10.1001/jama.2023.9618.
151. Interim measures for the management of generative artificial intelligence services. Beijing: Cyberspace Administration of China; 13 2023. (http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm, accessed 29 July 2023).
152. Satariano A. E.U. agrees on landmark artificial intelligence rules. *The New York Times*, 8 December 2023 (<https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html>, accessed 15 December 2023).
153. The EU should regulate on the basis of rights, not risks. Access Now, 17 February 2021 (<https://www.accessnow.org/eu-regulation-ai-risk-based-approach/>, accessed 21 June 2023).
154. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance *JAMA*. 2023;330(4):309–10. doi: 10.1001/jama.2023.9458.
155. Marcus G. Two models of AI oversight – and how things could go deeply wrong. Substack, 8 June 2023 (<https://garymarcus.substack.com/p/two-models-of-ai-oversight-and-how>, accessed 17 June 2023).
156. Kang C, Metz C. FTC opens investigation into Chat GPT maker over technology’s potential harms. *The New York Times*, 13 July 2023 (<https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html>, accessed 29 July 2023).
157. Ordish J. Large language models and software as a medical device. MedRegs blogs, 3 March 2023 (<https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/>, accessed 19 June 2023).
158. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med*. 2023. doi:10.1038/s41591-023-02412-6.

159. Ghost in the machine: Addressing the harm of generative AI. Forbrukerradet. Oslo: Norwegian Consumer Council; 2023 (<https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf>, accessed 9 July 2023).
160. Mökander J, Floridi L. Ethics-based auditing to develop trustworthy AI. *Minds & Machines*, 2021. doi:10.1007/s11023-021-09557-8.
161. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA*. 2023;330(4):315–6. doi: 10.1001/jama.2023.9651.
162. Questions and Answers: AI Liability Directive. Brussels: European Commission; 2022 (https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_5793, accessed 20 June 2023).
163. Duffourc MN, Gerke S. The proposed EU directives for AI liability leave worrying gaps likely to impact medical AI. *NPJ Digit Med*. 2023;6(1):77. doi:10.1038/s41746-023-00823-w.
164. Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023 (<https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf?sequence=1&isAllowed=y>, accessed 16 November 2023).
165. Marcus G. Artificial Intelligence is stuck. Here’s how to move it forward. *The New York Times*, 29 July 2017 (<https://www.nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html>, accessed 20 June 2023).
166. Parker G. Rishi Sunak to lobby Joe Biden for UK “leadership” role in AI development. *Financial Times*, 5 June 2023 (<https://www.ft.com/content/7c30ea28-2895-44c2-9a2d-c31ea7fa27e7>, accessed 19 June 2023).
167. Hogarth I. We must slow down the race to god-like AI. *Financial Times*, 13 April 2023 (<https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2>, accessed 10 July 2023).
168. Blinken A, Raimondo G. To shape the future of AI, we must act quickly. *Financial Times*, 24 July 2023 (<https://www.ft.com/content/eea999db-3441-45e1-a567-19dfa958dc8f>, accessed 30 July 2023).
169. Guterres, Antonio, Networked, inclusive multilateralism can help overcome challenges of era, says Secretary General, opening general assembly session, United Nations, 17 September 2019. (<https://press.un.org/en/2019/sgsm19746.doc.htm>, accessed 18 September 2023).