ISGlobal Barcelona Institute for Global Health

#ISGlobal_policy

A Compass to Navigate the Societal AI Transition and Advance AI for Global Health:

Opportunities, Challenges and Risks

An ISGlobal policy paper

Miguel Luengo-Oroz

JULY 2025

INDEX

The author would like to thank **Claudia García-Vaz**, **Quique Bassat**, and **Gonzalo Fanjul** for their contributions to the different drafts of this paper. The author also benefited from discussions in a workshop with researchers from IS Global. The author also wants to thank for their ideas and fruitful conversations researchers from the Spotlab.ai team.

This paper was drafted in 2024; therefore, the state of the field—including some concepts, technologies, and references—may have evolved since then, given the rapid pace of AI development.

With the support of:

aecid
Agencia Española de Cooperación Internacional para el Desarrollo

EXECUTIVE SUMMARY

Artificial intelligence (AI) is profoundly transforming various sectors, including global health, generating both high expectations and significant uncertainty. AI refers to machine-based systems designed to operate autonomously, producing outputs based on the inputs they receive. The deep learning approach that started in the 2010s has led to algorithms learning from examples and improving —in some cases, outperforming human capabilities. Governments are now developing strategies and regulations to manage its vast potential. The arrival of generative AI, which can create new content, from music to code, has sparked a debate about intellectual property. This technology has the potential to redefine innovation and creation much like the steam engine revolutionised manual labour. However, it also entails challenges, especially in labour market transitions, democracy, the impact of *deepfakes* and the societal value of trust in an already fragmented society, as well as the environmental and social footprint, a growing concern among environmentalists calling for more transparency. Health applications of AI range from the molecular to the societal level, enabling clinical applications as well. With promising results about the "humanity" touch of the interactions, but also resistance from healthcare workers.

In 2021, the UNESCO "Recommendation on the Ethics of Artificial Intelligence" was adopted by all 193 Member States. Its six core principles are protecting autonomy, promoting human well-being and safety, ensuring transparency, *explainability* and intelligibility, fostering responsibility and accountability, ensuring inclusiveness and equity and promoting responsive and sustainable AI. Sustainability and solidarity have initially been overlooked but are now gaining momentum, with the carbon footprint of AI systems and their environmental implications being now taken into account. Solidarity is slowly grabbing attention from an international law perspective.

Since 2017, countries have been developing national AI strategies that reflect their values and recognise AI's geopolitical importance. These strategies are shifting from advisory principles to enforceable regulations. The EU's AI Act introduces risk-based rules, banning harmful uses like behavioural manipulation or social scoring, and imposing strict requirements on high-risk applications. A key challenge is balancing regulation with innovation, especially compared to more flexible approaches in the U.S., U.K., and China. The rise of general-purpose generative AI (e.g., ChatGPT) has added complexity to regulation. These models must meet transparency standards, such as clear labelling of AI-modified or generated content and preventing illegal outputs. Models posing systemic risks require thorough evaluations and incident reporting, as will be the case for those in the health space. They will have to comply with medical device regulations as well as with other general AI regulations and privacy laws. Governments should work together in the development of international rules for the governance of AI, including all stakeholders and not just big tech firms.

The short-term risks arising from the implementation of AI health systems include bias, as they can inherit or exacerbate biases present in their training data, lack of transparency, the generation of convincing but fictional information, privacy violations of personal data and misuse for malicious purposes.

While the AI's potential to transform health and healthcare is huge, the path forward is full of challenges that must be anticipated and navigated carefully. It must be ensured that AI benefits all of humanity. Health research institutions have the chance to take a leadership role in this transition, but this would require a bold commitment at multiple levels.

SECTION 1.

Introduction

Artificial intelligence (AI) is changing every aspect of our society. It is impacting all geographies and sectors: economy —the highest valued companies in the world are devoted to working on AI—, society —the labour market is being changed—, and even war —with autonomous lethal weapons like kamikaze drones. In the field of global health, the conversation is well advanced and expectations are high. According to WHO Director General, Dr. Tedros Ghebreyesus *"AI is already playing a role in diagnosis and clinical care, drug development, disease surveillance, outbreak response, and health systems management [...] The future of healthcare is digital, and we must do what we can to promote universal access to these innovations and prevent them from becoming another driver for inequity."*[1]

In 2023, Ipsos ran a survey [2] on global attitudes towards AI, where two thirds of respondents anticipated AI will greatly change their lives in the near future. As we navigate through high levels of uncertainty, around half of respondents (54%) believe the benefits of AI surpass its drawbacks, while half of them feel AI products make them nervous. Interestingly, only 39% of respondents feel AI will benefit their health.

Whatever the future brings, global health research institutions will also be existentially impacted and must have a plan to adapt to the AI revolution, not only to navigate upcoming challenges but also to leverage its opportunities and maintain its relevance.

SECTION 2.

What Is AI and Where Is It Taking Us?

"The new AI systems present long term systemic challenges that are part of a bigger picture".

The European Union (EU) defines an "AI system" as *"a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content,* recommendations, or decisions *that can influence physical or virtual* environments".[3]

From a historical perspective, AI was first defined in 1955 during a summer school [4] where a group of scientists including Marvin Minsky and Claude Shannon considered that it would be interesting to explore how machines could simulate human thinking —a topic that turned out to be much more extensive than the two months that was supposed to last the summer school.

In its early stages, AI operated on **basic logical principles**, responding to direct stimuli with programmed responses. The approach was akin to playing chess: a different move prompts a specific response. This rule-based AI persisted until the paradigm shifted in the 1980s, with the emergence of **machine learning**, which applied statistical methods to decision-making processes.

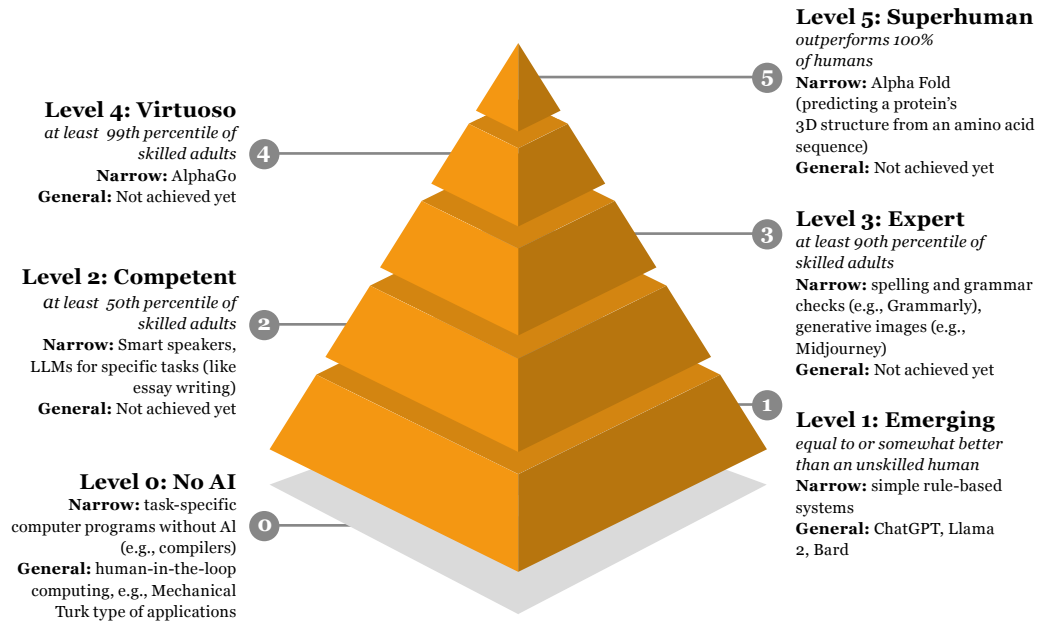
In the 2010s, the application and evolution of an old concept —artificial neural networks [5] —, in a new world full of data and exponential increases of computational power led to performance breakthroughs driven by this **deep learning approach**.[6] Artificial neural networks are computational models inspired by the human brain's structure, composed of layers of interconnected units (or 'neurons') that process information. Deep learning refers to using many of these layers to automatically learn complex patterns from large amounts of data. This led to algorithms learning from examples and performing better as they were exposed to more data. An illustrative example of this is image categorization: by feeding the AI numerous pictures of cats and dogs, it learns to identify and differentiate between the two. However, the processes behind this are based on massive statistical pattern recognition operators, not driven by an internal mental model. This means the AI might still confuse visually similar items, such as a chihuahua and a muffin (see Figure 1), due to its reliance on pixel recognition. In recent years, AI has not only matched, but in some cases **surpassed human capability** regarding image and speech recognition, as well as language translation.[7] AI's capacity to automate cognitive tasks at scale represents a critical general-purpose technology capable of having a global impact. Governments around the world, understanding its geostrategic value, have begun to formulate strategies and regulations to maximize its potential while addressing foreseeable risks of misuse.

FIGURE 1. AI program incapable of identifying between 2 similar sets of images.

Source: Teenybiscuit (2024). How to confuse Machine Learning. Twitter.

The emergence of **generative AI** represents a new *tour-de-force*. These systems, including algorithms capable of generating relatively original content across various domains —from text to images and beyond— are transforming content creation, design, and human-AI interaction.[8] They can create music, craft narratives, and produce code, reshaping how we approach creative and technical processes. While strong debates are still being held regarding intellectual property, since these models feed from original content from human artists and creators, technology keeps moving faster than we, as a society, are able to grasp. As generative AI evolves, it has the potential to redefine the boundaries of innovation and creation and give humans new abilities: a transition comparable to the impact the steam engine had on physical work.

FIGURE 2. Different levels of AI according to different capabilities.



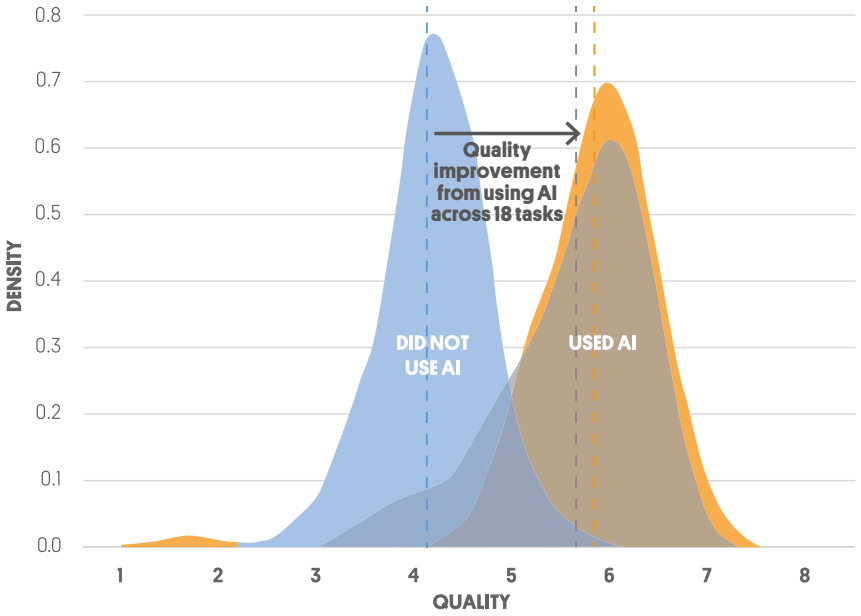
According to Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv:2311.02462 (2023), Google DeepMind
Center for Deep Tech Innovation

Source: Maslej N, Fattorini L, Perrault R, Parli V, Reuel A, Brynjolfsson E, et al. Artificial Intelligence Index Report 2024. arXiv (Cornell University). 2024; Available from: http://arxiv.org/abs/2405.19522.

These new AI systems present **long term systemic challenges** that are part of a bigger picture. These are some of the critical areas at stake:

- **Labour market transition:** Several studies show that AI enables workers to complete cognitive tasks more quickly and to improve the quality of their output *(see figure 3)*.[9] These studies also showed AI's potential to bridge the skill gap between low and high-skilled workers. The International Labour Organization estimates that 5% of the jobs will be automated with AI and – at least - 15% of the jobs will evolve significantly, with human capacities being augmented by AI.[10] It is imperative to think about transition plans for jobs so nobody is left behind, including healthcare workers.[11]

FIGURE 3. Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality.



Source: Morris MR, Sohl-Dickstein J, Fiedel N, Warkentin T, Dafoe A, Faust A, et al. Levels of AGI for operationalizing progress on the path to AGI. arXiv.org. 2023. Available from: https://arxiv.org/abs/2311.02462.

- **The impact on democracy:** Throughout the year 2024, more than two billion people casted their votes in over 50 countries. Generative AI tools are akin to social media on steroids, and we as a society are well aware of the impact social media has had on past elections. We are likely to witness new tactics to scrape up votes: imagine receiving personalized voice messages from a candidate directly to your WhatsApp. The landscape will be further complicated by both *deepfakes* [12] spreading falsehoods and uncomfortable truths misrepresented as *deepfakes*. In places where democracy and institutions are fragile, AI could be used to oppress minorities and their political candidates. To ensure democratic values are aligned with technological goals, every stakeholder, from technology developers and manufacturers to content disseminators and consumers, must bear responsibility and advocate for accountability mechanisms. These risks have been made ever more apparent as the technical progress of AI is led in non-democratic contexts (China) or by politically motivated actors.[13]

- **The societal value of trust:** Both the World Economic Forum and the United Nations (UN) have stated that 2024 was the year to rebuild trust.[14] Universal issues that seem to belong to no one, like climate change and misinformation, demand global cooperation founded on trust. Human societies are implicitly built on trust, yet what we are witnessing is a growing fragmentation of societal trust fuelled by the digital ecosystem. We are growing apart, divided into smaller, more polarized groups at every level, from international politics to local neighbourhood associations. A burning world full of AI-amplified narratives and a low threshold for war is a fertile ground for the emergence of even more divergent factions. What we need is, at the very least, a shared representation of the world. Although they are also very vulnerable to disinformation, science and the evidence-based scientific method could provide a robust foundation for building trust and ensuring that humanity retains common ground.[15]
- **Environmental and social footprint:** AI systems' deployment requires a lot of energy, leading to substantial carbon emissions which are still far from being the focus of the main conversation, but are a growing concern among environmentalists. In particular, training large AI models demands vast computational resources (the "cloud/computer farms") resulting in a large CO2 footprint,[16] and it requires use of rare minerals for electronic components,[17] which are often obtained in highly vulnerable contexts such as the DR of Congo. This environmental and social impact is further accentuated by the energy-intensive cooling systems necessary to prevent overheating in data centers, resulting in an important water footprint. It is urgent to implement sustainability policies, starting with reporting power and water requirements, as the AI industry scales up exponentially. As we navigate the climate emergency, AI should not be given a *laissez-passer*, and its purpose, justification and environmental impact must be carefully considered. Some promising initiatives advocate for an energy labelling system for AI models, along with the adoption of metrics such as 'tokens per watt per dollar'—where a token represents a unit of processed data, allowing for more transparent assessments of computational efficiency relative to energy consumption and cost.
- **Other AI existential risks (and black swans):** For some people, the potential of AI poses significant threats to human existence. Last year, a letter signed by a number of AI field leaders called for a moratorium on advanced AI systems development until their safety can be guaranteed and their risks managed.[18] Some of the strategies aimed at limiting the existential risks posed by superintelligent AI systems include preventing them from manipulating humans, disconnecting them from the internet and from robots, and not teaching them to program, thus preventing self-improvement. The main concern is that, in the future, superintelligent AI systems operating beyond our understanding could optimize their objectives in ways not aligned with human existence and human values. While such a scenario might seem unlikely, it is possible that we could first experience a different kind of shock concerning less intelligent AI systems being manipulated by malevolent actors. These "bad actors" could potentially attack energy networks, financial systems, and even engineer biological weapons.[19] We need robust AI safety measures and strategies and real-time surveillance mechanisms to prevent this, which requires all stakeholders to aim their efforts at managing and mitigating these risks.

BOX 1. Time for "Made in Human"

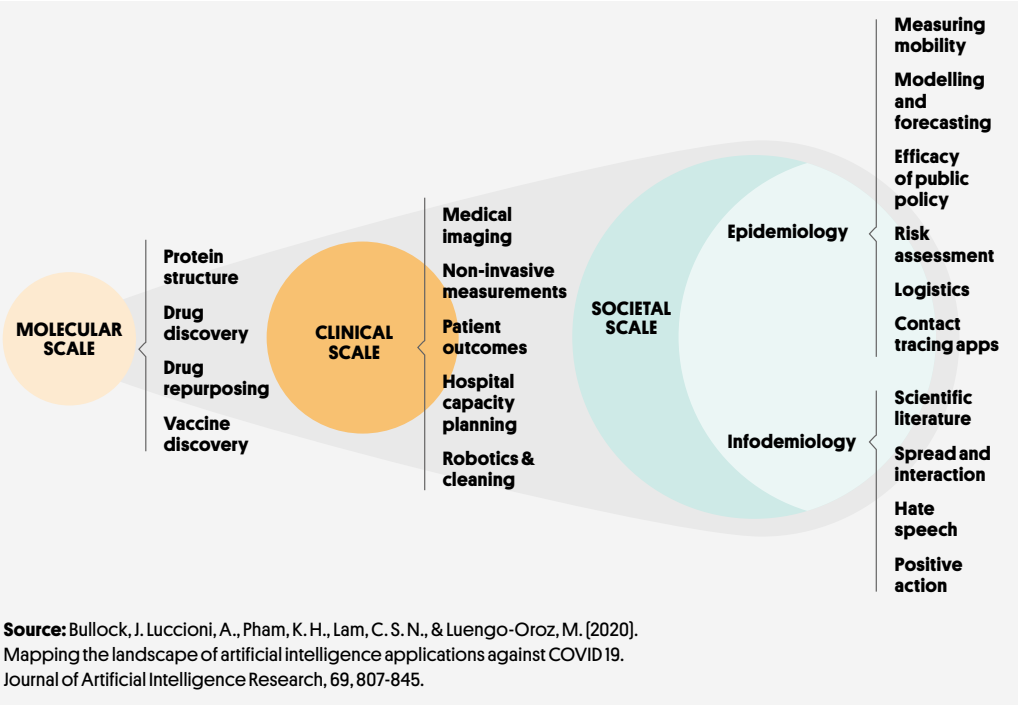
In the world of ideas, once only inhabited by human creations, now there are concepts created by AI or by humans using AI. With the explosion of synthetic data, we may soon live in a time where human content, the "Made in Human", becomes the exception.[20] This will require managing the flooding of artificial data. Assuming we establish the right rules and control the spread of synthetic content in our societies, a possible future approach might be to voluntarily watermarking and highlighting what's made by humans. Just as handmade items— like clothing or crafts—hold their value and are different from industrial goods, we could start to see human-made creations as uniquely special. The perfect human imperfection.

SECTION 3.

Which Are the AI Applications for Health?

To characterise AI applications in the field of health, we can look at different scales: [21] molecular, clinical, and societal *(see Figure 4)*.

FIGURE 4. Different AI use in health-related issues depending on its scale.



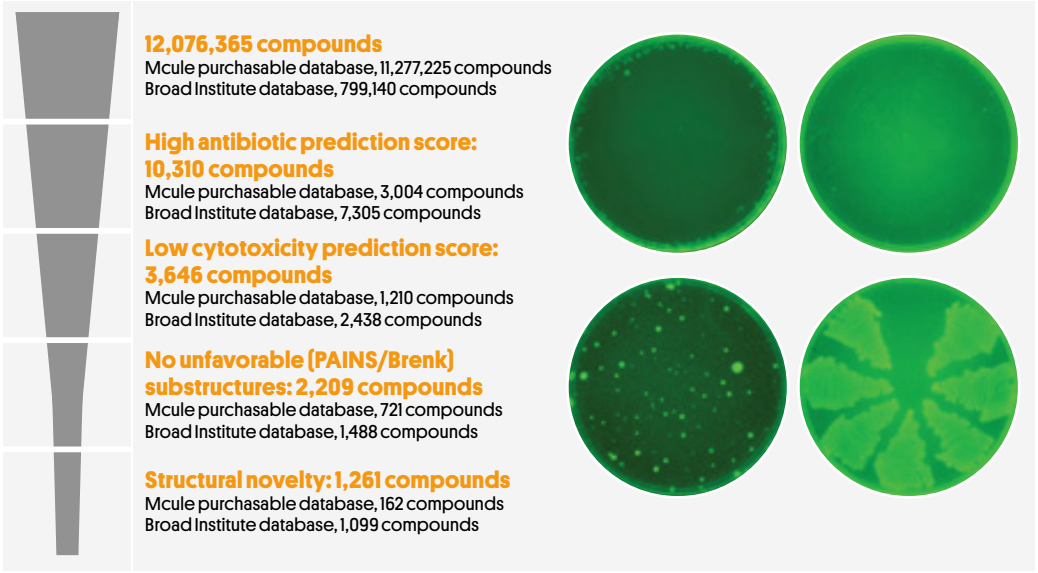
Source: Bullock, J. Luccioni, A., Pham, K. H., Lam, C. S. N., & Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID 19. Journal of Artificial Intelligence Research, 69, 807-845.

Source: Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence. 2022;4(3):189–91. Available from: https://doi.org/10.1038/s42256-022-00465-9.

"Clinical applications aim to enhance patient care from diagnosis through treatment and follow-up, and even predicting patient outcomes".

Applications at a **molecular** level deepen our understanding of biology, allow us to predict protein structures and functions,[22] repurpose existing drugs for new therapeutic uses, and discover new drugs.[23] Generative AI and the possibility of using large language models for genetics, instead of human language, open up unprecedented paths to accelerate drug and vaccine discovery and design. Recent research showed how an AI assisted funnel (a methodology that employs artificial intelligence to efficiently filter and analyze data) allowed to discover a new family of antibiotics[24] *(see Figure 5)*.

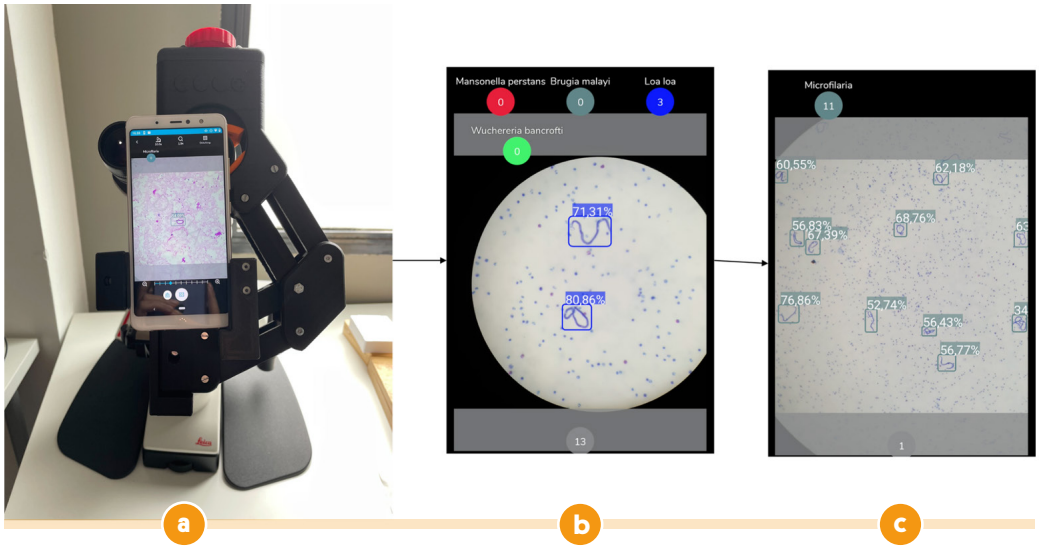
FIGURE 5. AI used to discover a new family of antibiotics.



Source: Vert JP. How will generative AI disrupt data science in drug discovery? Nature Biotechnology. 2023;41(6):750–1. Available from: https://doi.org/10.1038/s41587-023-01789-6.

Clinical applications aim to enhance patient care from diagnosis through treatment and follow-up, and even predicting patient outcomes. AI-driven tools, such as image-based diagnostics from various data modalities[25] and human-in-the-loop AI systems (an approach in which humans actively participate in AI decision-making and learning process) support medical professionals by reducing the time required to interpret data, as well as standardizing processes and measurements. For example, AI models running on smartphones can help diagnose neglected tropical diseases as filariasis from microscopy images in resource limited settings[26] *(see Figure 6)*. Wearables and other mobile devices allow for continuous health monitoring,[27] potentially tracking health trends and symptoms remotely, which saves healthcare resources and improves patient management. Integrating diverse data types, including all the information available in electronic health records can help creating AI driven subgroup stratifications,[28] risk indexes and individualized predictions, therefore allowing to forecast patient outcomes and resource needs. Finally, AI robotics is playing an increasingly essential role in routine tasks like the sanitization of healthcare facilities, as well as more refined ones like telemedicine or surgery.[29]

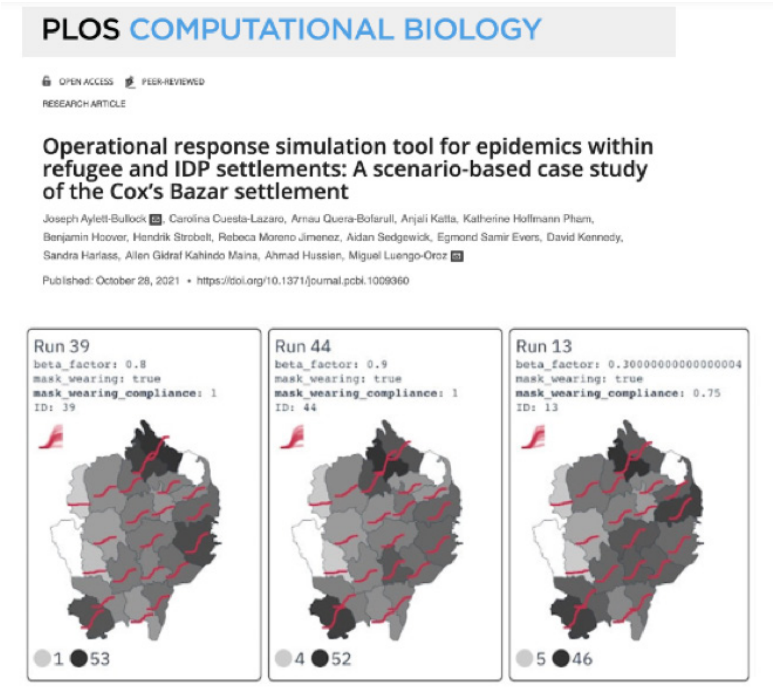
FIGURE 6. Edge AI for real-time automatic quantification of filariasis in mobile microscopy.



Source: Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nature Medicine. 2022;28(9):1773–84. Available from: https://doi.org/10.1038/s41591-022-01981-2.

On a **societal level**, AI can contribute to public health by improving early warning systems and allowing for public health intervention simulation. AI enhances traditional epidemiological methods by providing new nonlinear ways to analyze disease dynamics using multi-agent simulations (a computational model that involves multiple autonomous agents interacting within a defined environment) and digital twins (precise virtual representations of real-world physical objects, systems, or processes).[30] As an example, during the COVID-19 pandemic it was possible to create a digital twin model of the largest refugee camp in the world where more than 700 000 AI agents mimicking demographics and human behavior interact with each other simulating different public health interventions *(see Figure 7)*.[31] In addition to early warning and response, AI can support preparedness with a wide range of applications including forecasting diseases dynamics, resource allocation and societal behavior analysis.[32]

FIGURE 7. Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox´s Bazar settlement.



Source: Aylett-Bullock J, Cuesta-Lazaro C, Quera-Bofarull A, Katta A, Pham KH, Hoover B, et al. Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox's Bazar settlement. PLoS Computational Biology. 2021;17(10):e1009360. Available from: https://doi.org/10.1371/journal.pcbi.1009360.

At societal level we also see applications around *infodemics*.[33] AI not only makes it easier to produce misinformation and disinformation, but can also help fight it, clarifying and verifying content amid an overload of available data. AI assists in analysing audience engagement across platforms like social media,[34] aiding in rapid fact-checking and providing insights addressing health issues such as vaccine hesitancy. Research also focuses on monitoring and mitigating harmful content, such as hate speech and discrimination, which could hinder access to healthcare services. The chatbot revolution by Large Language Models (LLMs) makes a new wave of chatbots instrumental in distributing (hopefully) accurate and up-to-date health information widely and efficiently.

AI models also offer a new way of knowledge management, allowing scientists to speed up their research and summarise new knowledge for their own easy and fast comprehension, which can also be used for health science education purposes. Generative AI can also help streamline healthcare bureaucracy by, for instance, automating paperwork and filling out clinical documentation, drafting clinical visit summaries and, perhaps, making keyboards disappear as data from patient conversations would be automatically uploaded to the electronic health record with voice recognition analysis systems.[35] This would reduce time spent by healthcare providers in front of a computer and on administrative tasks, enhancing the overall efficiency of the healthcare system.

The rapid expansion of AI applications in healthcare emerges from a deep science perspective, but also from a profoundly human angle. A recent study showed that users perceived a *chatbot* designed to emulate doctors as more empathetic than actual doctors.[36] A different project found that nurses rated a nurse *chatbot's* empathy at the same level as their fellow human nurses.[37] Will machines be more empathic than humans? Certainly, they do not get tired or ever have a bad day, but this remains as a valid and open question. Perhaps the key is not to consider only the dialogue coming from the screen, but also the human presence. In May 2024, nurses from across California made a first demonstration of its kind in front of one major hospitals in San Francisco with their message: *"Trust nurses, not AI"*.[38]

FIGURE 8. Group of nurses protesting against the use of AI in one of San Francisco´s biggest hospitals.[37]

Source: The San Francisco Standard (2024). 'Trust nurses, not AI': Workers protest use of artificial intelligence at Kaiser hospitals.

BOX 2. A vision for AI in health

While it is hard to predict and imagine the future that these new cognitive superpowers offer to humanity, these are some of the visions on **how AI will accelerate science and healthcare delivery**:

- Decode biology and design new vaccines and medications.
- Eradicate diagnostic errors & delays
- Find the right personalized treatment and stewardship for every patient
- Accurate healthcare information for everyone everywhere in real time.
- Educate a new generation of healthcare workers with cognitive superpowers.
- Reduce health inequity by providing access to care for underserved communities.
- Design best evidence-based locally tailored public health interventions.
- Massive simulations to predict and influence One Health trajectories.

SECTION 4.

Which Are the Key Ethical Principles for AI Development and Use, and What Are the Governance and Regulatory Landscapes?

"Since 2017, countries have been developing their national AI strategies, which reflect their distinct values and priorities and demonstrate AI's geopolitical and geostrategic value".

Over the last decade, a myriad of principles and guidelines for AI development and use have emerged,[39] addressing both general and specific applications such as healthcare. The UNESCO inaugural *'Recommendation on the Ethics of Artificial Intelligence'* was adopted by all 193 Member States in November 2021.[40] The values and principles contained in the recommendation should be respected by all actors within the AI system life cycle, in the first place and, where needed, be promoted through development of new legislation, regulations and business guidelines. The values from the recommendation represent a common baseline for Member States and —with limitations due to the need for consensus— play a role as encouraging ideals in shaping policies. These include: (i) Respect, protection and promotion of human rights and fundamental freedoms and human dignity, (ii) recognize, protect and promote environment and ecosystem flourishing, (iii) ensuring diversity and inclusiveness and (iv) AI that enables peaceful and just societies, which is based on an interconnected future for the benefit of all. The principles —such as transparency, fairness, and human oversight— unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

TABLE 1. The six core AI principles approved by the WHO.

1	**Protecting autonomy:** AI should ensure that patients' autonomy and privacy are respected, and informed consent is maintained. This principle emphasizes the importance of AI transparency and the ability for individuals to make informed decisions about their own health (data).
2	**Promoting human well-being and safety**, aligning with SDG3: AI applications must prioritize the well-being and safety of individuals and public interest including robust measures to prevent harm and protect public health.
3	**Ensuring transparency, explainability, and intelligibility:** AI systems should be transparent and their operations explainable to ensure trust- so the design and functioning of AI should be clear to users and related stakeholders.

4 **Fostering responsibility and accountability:** Developers and users of AI systems must be responsible and accountable for their use. This includes adhering to legal and ethical standards as well as being prepared to address any negative impacts that arise.

5 **Ensuring inclusiveness and equity:** AI should be designed and implemented in ways that promote inclusiveness and equity, avoiding biases and ensuring that benefits and opportunities are distributed fairly across different populations, particularly the most vulnerable.

6 **Promoting responsive and sustainable AI:** AI must be continuously assessed to ensure they meet expectations and adapt to their specific contexts. Considering the AI environmental impact and ensuring that resources are used efficiently and ethically. particularly the most vulnerable.

Source: WHO. Health Ethics & Governance (HEG). Ethics and governance of artificial intelligence for health. 2021. Available from: https://www.who.int/publications/i/item/9789240029200.

Notably, two key principles initially overlooked in most recommendations are **sustainability** —taking into account the carbon footprint of AI systems and their environmental implications— and **solidarity** [41] —ensuring the prosperity generated by AI is shared, productivity gains are redistributed equitably, and AI does not exacerbate inequality while assessing the long-term implications before developing and deploying AI systems. Sustainability as a principle is gaining momentum, as the enormous carbon footprint of AI systems is being grasped.[42] Solidarity as a principle is slowly grabbing attention from an international law perspective, and recently the Independent Expert on human rights and international solidarity from the UN Human Rights issued a call for input regarding the upcoming report on Artificial Intelligence and international solidarity to the UN General Assembly.

The challenge of AI governance

In parallel to discussions around ethical principles and guidelines, since 2017, countries have been developing their national AI strategies,[43] which reflect their distinct values and priorities and demonstrate AI's geopolitical and geostrategic value. These strategies are evolving from principles —which are advisory and can be manipulated by lobbyists— to concrete regulations. The EU has recently approved the AI Act,[44] which aims, as the General Data Protection Regulation (GDPR) did, to set strict boundaries and procedures for human-centric AI. This risk-based regulation bans certain uses and requires high-risk applications to adhere to certified quality standards, while non-risk use cases are not subject to oversight. Unacceptable risk AI systems which are banned by the EU act include behavioral manipulation of people -for example AI toys that encourage dangerous behavior in children-, and social scoring systems based on socio-economic status or individual characteristics. A significant challenge for the EU will be creating a highly regulated environment which still fosters AI innovation, especially in comparison to the promotion of AI in the United States, United Kingdom or China.

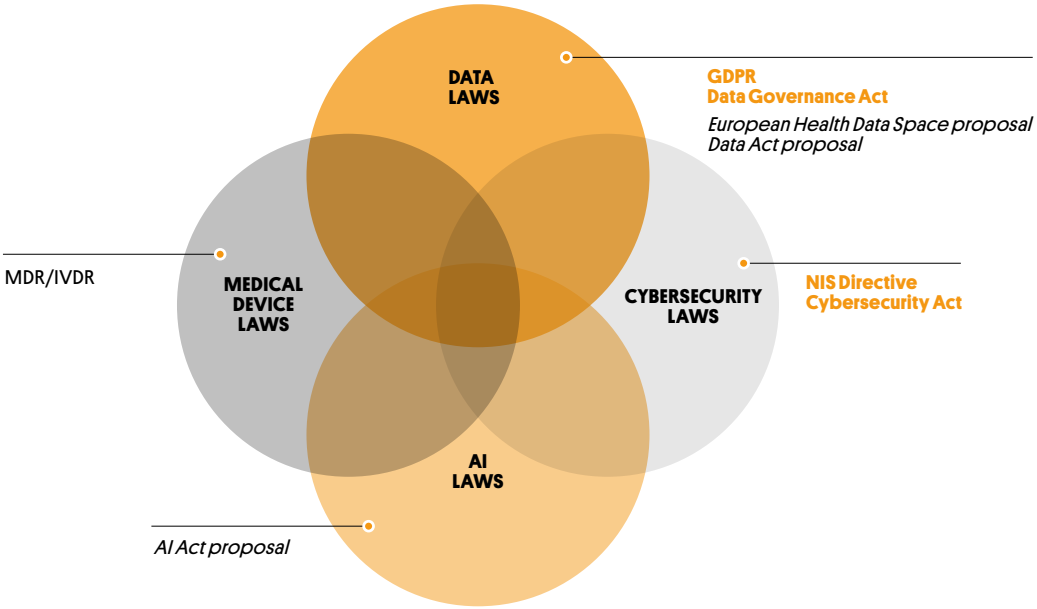
The arrival of general purpose generative AI models like *ChatGPT* has opened new fronts in the regulatory discussions. These models, for instance, are not considered high-risk under EU regulations but must adhere to specific transparency requirements. These include clearly disclosing AI-generated content, designing models to avoid producing illegal content, and publishing summaries of copyrighted data used in their training. Advanced general generative AI models which could present systemic risks, must undergo detailed evaluations, and any significant incidents will need to be reported to the EU. Additionally, any AI-generated or modified content, including images, audio, or video files like deepfakes, must be clearly labeled as AI-generated to ensure user awareness- this last point might be critical when building a future with trust in the digital world.

For those in the health space, most AI use cases will be considered high-risk and will need to comply with hard regulations. The regulatory landscape for healthcare AI intersects with at least four different areas which are moving target:[45]

- Medical device regulations: Ensure that healthcare technologies are safe and effective for patient use and AI systems used in medical settings fall under this category. In the EU, the Medical Device Regulation (MDR) and In Vitro Diagnostic Regulation (IVDR) are the key frameworks while in the US, the FDA oversees the regulation. AI technologies classified as medical devices must undergo rigorous testing, clinical evaluations, and continuous monitoring post-market to ensure safety and efficacy.
- General AI regulations (EU AI Act): The EU AI Act creates a regulatory framework for AI technologies, including those used in healthcare classifying the AI systems based on their potential risk, imposing stricter requirements on high-risk applications -such as healthcare AI. While the operational details of the AI Act are still to be defined, most of the requirements will align with the existing medical device regulation. Healthcare applications that did not fall into medical device categories will, however, have to follow new higher quality and regulatory standards than until now.
- Privacy laws (GDPR): Since healthcare AI systems process large amounts of super sensitive personal data, compliance with GDPR is a key requirement for any AI system, ensuring that data are handled legally and ethically. The arrival of generative AI is posing new challenges *vis-à-vis* privacy since the statistical nature of the models and the difficulty to delete data make it difficult to implement the rights of the individual over their data – as the right to be forgotten.
- Cybersecurity standards: Many of the critical risks stemming from AI systems have to do with bad actors. Ensuring compliance with cybersecurity standards can be challenging due to the rapidly evolving nature of cyber threats- as AI also provides new ways to attack healthcare systems. AI for health providers shall try to follow international standards such as the ISO/IEC 27001 which outline best practices for information security management.

Beyond the EU, other regions are currently defining their regulatory approaches; for example, the Africa CDC is beginning to establish AI regulations for its members. However, it's still early days for most, and stakeholders looking to deploy AI in healthcare will need to continually adjust their strategies and operations.

FIGURE 9. New cybersecurity requirements for medical devices in the EU: the forthcoming European Health Data Space, Data Act, and Artificial Intelligent Act.



Source: Mirsky Y, Mahler T, Shelef I, Elovici Y. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. arXiv.org. 2019. Available from: https://arxiv.org/abs/1901.03597.

Overall, it is clear that governments should work in the collective development of international rules for the governance of AI. Among the initiatives working in the subject, the UN's High-level Advisory Body on AI [46] is set to undertake analysis and advance recommendations for the international governance of AI. As these international frameworks are developed, it is vital that everyone have a seat at the table —not just those from high-income countries, but also those from low— and middle-income countries, and not just big tech firms but all stakeholders including academia, and civil society —to ensure a multitude of voices is heard and can contribute to the direction humanity will take with AI.

SECTION 5.

Which Are the Main Risks and Challenges for Trustworthy Health-Related AI?

There are a number of immediate and short-term risks arising from the implementation of (hopefully) trustworthy AI systems which include:

- **Bias:** AI systems can inherit or even exacerbate biases present in their training data, leading to unfair or prejudiced outcomes, particularly in sensitive domains. In practice, robust and standardized evaluations for responsible AI are lacking. As an example, an AI system used for screening of pulmonary diseases which was trained with data from white men from North America could fail in unknown ways when used in women from Southeast Asia if that demography has not been properly taken into account.
- **Explainability:** Many AI models, particularly those based on neural networks, can be seen as "black boxes," offering little insight into how they arrived at a given output. Though there is ongoing research along these lines, black box approaches make it difficult to audit and trust the model decisions, especially in high-stakes scenarios. As an example, a health insurance provider might be making decisions over coverage or pricing based on AI systems that classify clients without an explainable criterion.
- **Veracity and hallucinations:** AI systems sometimes generate convincing but entirely fictional information, known as "hallucinations." Ensuring the veracity of AI-generated content is a key and significant challenge for the adoption of medical chatbots. We could imagine a medical doctor that trusts too much its AI copilot - as it usually makes right advice- and promotes the wrong decision based on AI hallucinations. This has been particularly noticeable when referencing previous work in the literature, with entirely fictional bibliographical references used to back up information generated by AI.
- **Privacy violations:** AI systems can intrude on personal privacy by collecting, analyzing, and potentially mishandling personal data. Patient re-identification becomes easier with multi-modal data from patients. Furthermore, it is extremely complicated to make AI and neural networks forget.[47]
- **Accountability:** Lack of transparency in AI decision-making can obscure accountability, making it unclear who is responsible for AI's actions or mistakes. Technically, the probabilistic nature of large language models, the complexity of reproducing the AI models due to scale, and the lack of traceability requirements of AI models make it difficult to reconstruct accountability chains.
- **Misuse for nefarious purposes:** AI can be weaponized or used for malicious purposes such as privacy violation with deepfakes, automated cyberattacks on health systems, or mass surveillance. As an example, researchers have shown the possibility of creating deepfake MRI images showing non-existent brain tumours inserted in original images of healthy individuals.[48] Generative AI used to design new drugs can also be used to design chemical weapons.[18]

"AI systems can inherit or even exacerbate biases present in their training data, leading to unfair or prejudiced outcomes, particularly in sensitive domains".

Conclusion

As AI continues to revolutionize every sector, it is having a profound impact on scientific research and health, therefore changing global health, from research to public health interventions. This paper has explored the trajectories of AI from multiple perspectives. While the AI's potential to transform healthcare is huge—from molecular research to clinical applications and public health interventions—, the path forward is full of challenges that must be anticipated and navigated with care.

The key ethical principles and regulatory landscapes discussed highlight the necessity of a robust framework to govern AI development and deployment. Transparency, accountability, inclusivity, sustainability and solidarity are crucial to building trust and ensuring that AI benefits all of humanity. The intersection of AI with societal issues such as labour markets, democracy, trust, and environmental sustainability reinforces the need for comprehensive strategies to manage AI's impact.

Health research institutions such as ISGlobal stand at a critical juncture, as they, for sure, will have to adapt to remain relevant– and can also choose to lead the AI transformation in particular domains. This requires a bold commitment of the organization at multiple scales and across all areas of work, from education and research to policy and innovation.

In conclusion, the future roadmap for global health will be heavily shaped by how we navigate with concrete actions the AI transition in the short term and where we point the compass for the medium- and long-term vision. Institutions have now the opportunity to lead by example, driving responsible AI innovation that benefits all.

REFERENCES

1. WHO (2024). Harnessing Artificial Intelligence for Health.

2. Ipsos (2023). Global views on A.I. 2023. How people across the world feel about artificial intelligence and expect it will impact their life. July 2023.

3. European Parliament (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (Artificial Intelligence Act).

4. McCarthy J, Minsky ML, Rochester N, Shannon CE. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine. 2006;27(4):12.

5. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin Of Mathematical Biophysics. 1943;5(4):115-33. Available from: https://doi.org/10.1007/bf02478259

6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44. Available from: https://www.nature.com/articles/nature14539

7. Maslej N, Fattorini L, Perrault R, Parli V, Reuel A, Brynjolfsson E, et al. Artificial Intelligence Index Report 2024. arXiv (Cornell University). 2024; Available from: http://arxiv.org/abs/2405.19522

8. Morris MR, Sohl-Dickstein J, Fiedel N, Warkentin T, Dafoe A, Faust A, et al. Levels of AGI for operationalizing progress on the path to AGI. arXiv.org. 2023. Available from: https://arxiv.org/abs/2311.02462

9. Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN Electronic Journal. 2023; Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321

10. Generative AI and Jobs: A global analysis of potential effects on job quantity and quality. Available from: https://webapps.ilo.org/static/english/intserv/working-papers/wp096/index.html

11. Spatharou A, Hieronimus S, Jenkins J. Transforming healthcare with AI: The impact on the workforce and organisations. McKinsey & Company. 2020.

12. JUST JOKING! Deepfakes, Satire, and the Politics of Synthetic Media - MIT CoCreate [Internet]. MIT CoCreate. 2022. Available from: https://cocreationstudio.mit.edu/just-joking/

13. Schaake, M. (2024). *The Tech Coup: How to save democracy from Sillicon Valley*. Princeton University Press.

14. World Economic Forum. 2024. Davos 2024: Rebuilding trust in the future.

15. National Academies Press (2023). 2023 Nobel Prize Summit Truth, Trust, And Hope.

16. Luccioni S, Jernite Y, Strubell E. Power hungry processing: Watts driving the cost of AI deployment? 2022 ACM Conference on Fairness, Accountability, and Transparency. 2024; Available from: https://arxiv.org/abs/2311.16863

17. Crawford, K. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. 2021. Yale University Press. Available from: https://doi.org/10.2307/j.ctv1ghv45t

18. The New York Times. 2023. A.I. Poses 'Risk of Extinction', Industry Leaders Warn.

19. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence. 2022;4(3):189–91. Available from: https://doi.org/10.1038/s42256-022-00465-9

20. Made in Human. 2024. A framework to build trust and promote transparency in the era of Artificial Intelligence. Available from: https://www.madeinhuman.org/english

21. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M. Mapping the landscape of Artificial Intelligence applications against COVID-19. Journal of Artificial Intelligence Research. 2020;69:807–45. Available from: https://arxiv.org/abs/2003.11336

22. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

23. Vert JP. How will generative AI disrupt data science in drug discovery? Nature Biotechnology. 2023;41(6):750–1. Available from: https://doi.org/10.1038/s41587-023-01789-6

24. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2023;626(7997):177–85. Available from: https://www.nature.com/articles/s41586-023-06887-8

25. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nature Medicine. 2022;28(9):1773–84. Available from: https://doi.org/10.1038/s41591-022-01981-2

26. Lin L, Dacal E, Díez N, Carmona C, Ramirez AM, Argos LB, et al. Edge Artificial Intelligence (AI) for real-time automatic quantification of filariasis in mobile microscopy. PLoS Neglected Tropical Diseases. 2024;18(4):e0012117. Available from: https://doi.org/10.1371/journal.pntd.0012117

27. Smith AA, Li R, Tse ZTH. Reshaping healthcare with wearable biosensors. Scientific Reports. 2023;13(1). Available from: https://www.nature.com/articles/s41598-022-26951-z

28. Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. New England Journal of Medicine. 2023;389(13):1211–9. Available from: https://doi.org/10.1056/nejmra2212850

29. Kyrarini M, Lygerakis F, Rajavenkatanarayanan A, Sevastopoulos C, Nambiappan HR, Chaitanya KK, et al. A survey of robots in healthcare. Technologies. 2021;9(1):8. Available from: https://www.mdpi.com/2227-7080/9/1/8

30. Katsoulakis E, Wang Q, Wu H, Shahriyari L, Fletcher R, Liu J, et al. Digital twins for health: a scoping review. Npj Digital Medicine. 2024;7(1). Available from: https://www.nature.com/articles/s41746-024-01073-0

31. Aylett-Bullock J, Cuesta-Lazaro C, Quera-Bofarull A, Katta A, Pham KH, Hoover B, et al. Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox's Bazar settlement. PLoS Computational Biology. 2021;17(10):e1009360. Available from: https://doi.org/10.1371/journal.pcbi.1009360

32. Syrowatka A, Kuznetsova M, Alsubai A, Beckman AL, Bain PA, Craig KJT, et al. Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. Npj Digital Medicine. 2021;4(1). Available from: https://doi.org/10.1038/s41746-021-00459-8

33. Calleja N, AbdAllah A, Abad N, Ahmed N, Albarracin D, Altieri E, et al. A Public Health Research Agenda for Managing Infodemics: Methods and Results of the first WHO Infodemiology Conference. JMIR Infodemiology. 2021;1(1):e30979. Available from: https://doi.org/10.2196/30979

34. Ng YMM, Pham KH, Luengo-Oroz M. Exploring YouTube's recommendation system in the context of COVID-19 vaccines: Computational and comparative analysis of video trajectories. Journal of Medical Internet Research. 2023;25:e49061. Available from: https://doi.org/10.2196/49061

35. Topol, E (2019). Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK.

36. 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 Internal Medicine. 2023;183(6):589. Available from: https://doi.org/10.1001/jamainternmed.2023.1838

37. Hypocratic AI (2024). Safety focused generative AI for healthcare. Available from: https://www.hippocraticai.com/

38. The San Francisco Standard (2024). 'Trust nurses, not AI': Workers protest use of artificial intelligence at Kaiser hospitals.

39. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019;1(9):389–99. Available from: https://www.nature.com/articles/s42256-019-0088-2

40. UNESCO (2022). Recommendation on the Ethics of Artificial Intelligence. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000381137

41. Luengo-Oroz M. Solidarity should be a core ethical principle of AI. Nature Machine Intelligence. 2019;1(11):494. Available from: https://www.nature.com/articles/s42256-019-0115-3

42. Zomorodi M. ChatGPT vs. the climate: The hidden environmental costs of AI. NPR. 2024. Available from: https://www.npr.org/2024/05/10/1250261120/chatgpt-vs-the-climate-the-hidden-environmental-costs-of-ai

43. Dutton T. An Overview of National AI strategies - Tim Dutton - medium. Medium. 2024. Available from: https://medium.com/@tim.a.dutton/an-overview-of-national-ai-strategies-2a70ec6edfd

44. EU AI Act: first regulation on artificial intelligence | Topics | European Parliament. 2023. Available from: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

45. Biasin E, Yaşar B, Kamenjašević E. New cybersecurity requirements for medical devices in the EU: the forthcoming European Health Data Space, Data Act, and Artificial Intelligence Act. 2023.

46. United Nations (2023). Governing AI for Humanity. Advisory body on Artificial Intelligence. Available from: https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

47. Luengo-Oroz M. We forgot to give neural networks the ability to forget. Forbes. 2023. Available from: https://www.forbes.com/sites/ashoka/2023/01/25/we-forgot-to-give-neural-networks-the-ability-to-forget/

48. Mirsky Y, Mahler T, Shelef I, Elovici Y. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. arXiv.org. 2019. Available from: https://arxiv.org/abs/1901.03597

www.isglobal.org

@ISGLOBALorg
/isglobal
@ISGLOBALorg
/isglobalorg
@isglobal.org
@isglobalorg

ISGlobal Barcelona Institute for Global Health

A partnership of:

"la Caixa" Foundation

Clínic Barcelona
Universitat de Barcelona
Generalitat de Catalunya
Gobierno de España
Hospital del Mar Barcelona
upf. Universitat Pompeu Fabra Barcelona
Ajuntament de Barcelona