Research Paper

Digital Society Initiative

June 2024

Artificial intelligence and the challenge for global governance

Nine essays on achieving responsible AI

Alex Krasodomski (editor), Arthur Gwagwa, Brandon Jackson, Elliot Jones, Stacey King, Mira Lane, Micaela Mantegna, Thomas Schneider, Kathleen Siminyu and Alek Tarkowski

CHATHAM HOUSE

Chatham House, the Royal Institute of International Affairs, is a world-leading policy institute based in London. Our mission is to help governments and societies build a sustainably secure, prosperous and just world.

Contents

Foreword

When the UK hosted its AI Safety Summit on 1–2 November 2023, the country's prime minister, Rishi Sunak, used the occasion to interview the entrepreneur and chief of X, Tesla, SpaceX, Neuralink and The Boring Company, Elon Musk.

There was some trepidation about this. It was not just that Musk is a controversial figure, but that the prime minister – the democratically accountable leader of a G7 nation – was the interviewer rather than the interviewee, the questioner rather than the one with the answers. The tableau crystallized a shifting landscape – one where leaders of technology companies wield significant power, and where leaders of states seem to come to them for solutions.

The theorist Ian Bremmer and Mustafa Suleyman, the co-founder of DeepMind, argued in 2023 that we are living in a 'technopolar' world – where power is wielded not just through control of capital, territory or borders, but through control of computing capacity, algorithms and data.[1] Under this model, tech companies significantly shape how ordinary people interact with the world, and are similarly consequential for labour markets and geopolitics.

Nowhere was this more clearly underlined than in Ukraine in 2022–23. Musk's Starlink satellite internet services had emerged as a critical capability of the Ukrainian resistance, but their provision was dependent on a private company, leading to uncertainty over who called the shots on their use and availability.

If warfare is changing, so too are international norms: decisions affecting ordinary people in all sorts of ways are increasingly made in Silicon Valley boardrooms. Norms on privacy, access to information and freedom of expression are set out in terms of service for billions of digital platform and software users worldwide. Such consolidation makes perfect sense from the major tech firms' perspectives; after all, local laws, languages and values create costly administrative inefficiencies in globally minded businesses.

But this narrative of tech power also misses some of the critical ways in which states still shape tech companies' ability to act. Governments have not stayed on the sidelines in response to Big Tech's more prominent role, or potential role, in geopolitical events. Indian state authorities, for example, are known for the frequency with which they shut down internet access to avoid social or political unrest;[2] and India is increasingly shaping norms about the censorship of social

1 Bremmer, I. and Suleyman, M. (2023), 'The AI Power Paradox', *Foreign Affairs*, 16 August 2023, https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox.
2 Software Freedom Law Center India (2024), Internet shutdown tracker, https://internetshutdowns.in (accessed 11 April 2024).

media networks such as X (formerly Twitter).[3] The Chinese state has long sought to challenge Western hegemony in internet architecture, and to influence global digital governance standards. Export controls imposed by the Biden administration in the US affect where and how the tech industry locates factories and develops advanced chips. The concept of the tech company, or even the private sector, as entirely separate from the state is not a reality everywhere.

State power and tech power interact, and have long done so. The question facing us in the years to come is how those relationships may change or break down. At the fringes of technology – from artificial intelligence (AI) to quantum computing – state power can feel scarce. With some exceptions, when governments do come to the table, they arrive too late or too poorly staffed to be seen as equals: well-meaning bureaucrats at best, a handbrake on profit or progress at worst.

Chatham House is interested in posing questions about this – and, ideally, answering some of them. What would it take for co-governance of technology by the state and the private sector, and how can states around the world adapt to the rising power of tech companies, collaborate with them, and coordinate responses and regulation? What is the extent of Big Tech's power on policymaking today? States, after all, are politically answerable for many of the decisions affecting their citizens even where those decisions are currently made in boardrooms, not in parliaments or ministries. If, as the economist Mariana Mazzucato suggests, governments need to learn how to row the boat so they can steer it, then states need to learn how to build and make tech, not just interact with it, to steer their way through 21st-century challenges.

By turn, it is a moment for industry to look at itself, and ask whether it is able to deliver the public goods it sometimes touts, and how it can steward and respond to the consequences of vast technological change. How both sides broker the relationship between tech power and state power is going to shape geopolitics in the future.

These questions are especially acute in a year in which half the world is going to the polls. Some people are concerned about a future of electoral 'post-reality' shaped by AI-enabled mis- or disinformation. Our window onto politics and candidates will be framed by the technology that mediates our access to news and information. In this collection of essays on AI's implications for society and governance, and in our ongoing work at Chatham House, we explore these questions: looking at the merits of community-driven AI, unpacking the challenges around international cooperation and efforts to establish common rules, discussing AI 'decolonization', arguing the case for open-source AI development and more.

Bronwen Maddox
Director and chief executive, Chatham House

---

3 Mehrotra, K. and Menn, J. (2023), 'How India tamed Twitter and set a global standard for online censorship', *Washington Post*, 8 November 2023, https://www.washingtonpost.com/world/2023/11/08/india-twitter-online-censorship.

Summary

- Artificial intelligence (AI) is creating novel challenges for governance. Technical advances and widening use of AI have increased concerns about the risks of such technology, while also underscoring its transformational potential. This collection of nine essays explores pathways towards responsible AI, and proposes both broad principles and specific ideas for future-proof AI governance.
- The inadequate regulation of AI to date is hardly for lack of recognition of the risks. The meteoric rise of generative AI has dramatically raised AI's profile in the public debate. Amid wild predictions about what an AI-dominated future might look like, there have also been serious efforts to write new laws on AI. For instance, the EU's AI Act establishes protections around the use of biometric data in law enforcement contexts, and imposes restraints on the use of AI systems in high-risk applications like self-driving cars or healthcare.
- But regulation remains fragmented, with limited coordination or harmonization between jurisdictions. The new AI Act, for instance, is intended for the EU's single market but lacks legal force elsewhere. It also reflects a distinctive European approach at odds with regulatory cultures and political technology agendas in the US and China, the two other major hubs of AI development.
- More positively, the global governance gap is prompting policy innovation. AI safety institutes are attracting new talent into governance. The Council of Europe has developed an AI treaty that establishes a binding legal framework on human rights, democracy and the rule of law for AI. The treaty is expressly designed to provide a model for legislation beyond the Council of Europe's 46 members, so that other countries can develop 'differentiated' laws tailored to their own contexts. The framework format potentially offers a way of keeping pace with rapid technological changes by allowing laws to be developed continually in future, in accordance with the treaty's guiding principles. More broadly, systems supporting more agile and dynamic governance might eventually allow regulatory 'releases' to be published in smaller and faster steps, much as currently happens with software updates. AI systems themselves might even end up being used to regulate AI.
- An alternative way to coordinate global regulation could be to establish an international AI research facility modelled roughly on CERN, the particle accelerator and nuclear physics lab on the Franco-Swiss border. The idea would be to emulate CERN's spirit of international collaboration, enabling the pooling of resources beyond those available to individual countries. In particular, a CERN-type approach could support the very large computing infrastructure needed to process vast amounts of data for AI. It could stimulate public sector

funding, insulate research from national political agendas, redistribute talent away from private AI labs, and reduce the moral hazard potentially associated with giving private firms a leading role in shaping governance of the very technologies they create and sell.

— This latter risk underlines the tension between the public and private sectors that pervades almost all discussion of AI regulation. Put simply, large AI labs such as Anthropic, Google DeepMind and OpenAI have dominated the technical development of AI. The concern is that this could entrench a power imbalance between regulators and the regulated, in the latter's favour – governance approaches to date have not resolved this issue. Big tech firms have signalled their recognition of the potentially unprecedented risks from AI, and their own eagerness to assist with its responsible regulation. But it remains to be seen how effectively this commitment will translate into practice.

— One obvious way to respond to the risk of private capture of AI is to support a viable public sector alternative. This would not be easy, but there is a historical precedent. In the UK, the birth of the BBC in 1922 was a direct response to the rise of radio, at the time a revolutionary technology that worried the British establishment. This is where a so-called 'public option' for AI could come in. A publicly owned British AI Corporation (BAIC) – a kind of BBC for AI, as it were – could start to address the lack of popular trust in the state's ability to build reliable technology. It would need a charter and 'usage-based' financial model that ensured its independence and commercial sustainability.

— Widespread acceptance would rely on a BAIC developing AI applications of genuine public utility: tools that addressed concerns around privacy, job security and equality, for instance, rather than churning out superficial entertainments or amplifying misinformation. A BAIC could sustain creative sectors by paying for training datasets rather than following the common practice of 'scraping' the internet for data. Although the relevant essay in this collection considers the UK-specific possibilities of public-option AI, the model could be extended to many countries.

— Responsible AI may also mean protecting the principles of open-source development, as well as inclusivity, fairness and equality. In the past, embryonic forms of AI were largely built along open-source lines, but that trend has recently reversed. Partly due to professed concerns about security, the big AI industry players have increasingly moved from open-source models to closed, proprietary approaches. Critics contend that this is more about protecting market share and reducing competition than about improving safety.

- Any concentration of tech power has a number of potential drawbacks. It can result in the unequal distribution of the economic benefits of AI. It can also encourage an unhealthy *cultural* centralization and homogenization of technology around Western identities and values. In short, if tech firms in Silicon Valley and elsewhere in the West determine the trajectory of AI, then AI is more likely to reflect and cater to those dominant communities. Under-represented and marginalized communities, cultures and languages risk getting left behind. Many of the datasets used to train AIs are predominantly English-based, with the risk that the resultant tools work poorly in other languages, disregard the needs of non-anglophone cultures, and facilitate AI solutions that exclude or discriminate against those cultures. These failings can be characterized as a form of AI 'colonialism'.
- However, a more optimistic future is imaginable, in which universal rules on AI are jointly shaped in a global public sphere drawing on many cultures and value systems. Promising grassroots work is now being undertaken to democratize AI, and to give greater emphasis to non-Western cultures and languages. One example is BLOOM, a large language model released in 2022 that features open-source code, transparent training datasets and a collaborative production model. BLOOM works in 46 languages and is based on 'justly sourced' data. In Africa, the concept of community-based AI is taking off, as a thriving grassroots AI ecosystem creates new AI tools in African languages, offers educational opportunities to local coders, and promotes a non-Western vision of AI.
- AI's ethical implications also need clearer recognition in governance. The rollout of AI will embed automated decision-making in many areas of our lives. There is a critical need for ethical frameworks to determine what should be automated and what should not, and when human oversight of AI systems is essential. An AI that recommends which movie to watch does not need the same safeguards as one that makes a parole decision. The ethical challenges are compounded by the nature of generative AI, which can make it hard to distinguish convincing AI 'hallucinations' from reality. AI systems that perpetuate biases and prejudices render automated decisions unreliable, particularly if the datasets on which they are based are insufficiently representative of diversity.
- Ultimately, this essay collection illustrates the need for dynamic and multi-stakeholder approaches to governance. Responsible development of AI cannot occur in silos. It needs to be jointly and cooperatively guided, through global processes for reconciling competing interests and agreeing priorities.

01
Introduction – the need to future-proof AI governance

Artificial intelligence is changing so rapidly that its would-be regulators are having trouble keeping up. But the potential impacts of AI on societies may be so transformative – whether for better or worse – that strengthening cooperative, global governance to ensure a future of responsible AI is an urgent necessity.

Alex Krasodomski

In 2023, Chatham House asked its network of digital technology and policy experts for their big questions on artificial intelligence (AI). This essay collection – the inaugural paper in a planned series of publications on AI – sets out to offer some answers. Written by leaders in their fields, the essays present a range of perspectives on the promise and pitfalls of efforts to govern this emerging technology. The collection brings together voices from industry and government, civil society and academia, and perspectives from Africa, Europe, Latin America, the UK and the US.

Governance of emerging technologies such as AI may prove to be one of the defining challenges for international relations in the 21st century.[4] The competition for technological hegemony promises its winners economic advantage, the entrenchment of their values and norms, and an edge in military power. China and the US – locked in an increasingly tense rivalry in many areas, including technology – remain the most significant investors in AI development globally.

Yet if 20 years of digital technology development have proven one thing, it is that power derived through technology rarely maps neatly to geographies, markets or any existing set of international rules, norms or values. New centres of power have emerged. Governance of technology is usually retroactive. Institutions – whether democratic or autocratic, in politics, media and throughout the economy – need time to come to terms with the changing technology landscape and to adapt accordingly. But technology advances rapidly, which means that decisions made in corporate boardrooms often precede and have more weight than those made in parliaments,

4 Dafoe, A. (2018), *AI Governance: A Research Agenda*, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

government ministries or regulatory agencies. Whether intentionally or inadvertently, companies are often in effect setting global standards on fundamental rights, on political and social norms, and on the assumptions, aims and values that shape the technology we use in modern life.

While quick to spot the opportunities, national governments in particular have been slow to rise to these challenges, and multilateral institutions even slower. Shaping foundational digital technologies – digital media, sharing platforms, cloud storage, encrypted messenger apps and now AI – remains a point of weakness for most governments, particularly democracies.

Why does this matter now in particular? There is an emerging consensus that the stakes are higher than before as a result of this next wave of technologies. Even the most sceptical observer of AI development would agree that AI will be responsible for significant upheaval: it will certainly disrupt economies, societies and many dimensions of physical and digital security; the impacts will likely be even broader as the technology continues to be deployed more deeply into our everyday lives. AI prophets might go further. They might promise, or warn of, a reassessment of the most fundamental aspects of global society – encompassing understandings of economic value, questions around the superiority of humans to machines, or even the likely survival of humans as a species.

How AI will transform the world is a geopolitical question. Conflict and competition will shape the technology in certain ways; cooperation will shape it in others.

How AI will transform the world is a geopolitical question. Conflict and competition will shape the technology in certain ways; cooperation will shape it in others. AI designed and built in a cutthroat marketplace will look different to AI dominated and shaped by monopoly power. AI development led by universities will not resemble that led by states, militaries, philanthropic organizations or technology companies. AI developed in China will be different to AI built in the US, Europe or India. Which of these trajectories are more or less likely, whether some might coexist, and how they can be steered are the questions at the heart of this essay collection.

Taken as a whole, the authors' arguments on various dimensions of AI governance underscore the idea that ensuring collective human benefit should be the guiding principle for negotiations on the future of AI. Beyond a focus on risk aversion or harm prevention, the authors demand a clear articulation of the kind of world we should be aiming to bring about through this technology. Perhaps above all else, the collection is a call for clarity from those around the table about their aims, and about the realities of this technology revolution.

Whether change needs to come from the design of new institutions or regulatory frameworks, from multi-stakeholder consultation or from community leadership, the message is clear: current AI governance is insufficient. It is insufficiently

incentivized, insufficiently resourced, insufficiently coordinated and insufficiently representative. AI governance will need new agreements, treaties and institutions: a CERN-like institution for global cooperation on AI research, for instance, or new corporate models and multilateral treaties governing the use of AI.

Without a change, we risk repeating and entrenching blunders made in the provision of digital technology in recent decades. While the proponents of AI may be fond of emphasizing its novelty, there are deep and disconcerting continuities at the heart of the AI revolution. Access to digital technology remains wildly uneven around the world on any measure: internet connectivity, advertising spend, the availability of affordable mobile internet services, investment in digital infrastructure.[5] Improved access to AI is essential: through skills development, infrastructural investment and the thoughtful use of open-source approaches. The race for market share by US and Chinese technology firms is a familiar story that carries lessons for the next generation of AI-enabled technology, particularly when considering the often questionable effectiveness of regulation in anticipating future technical developments, and the insufficient influence of global majority countries on the technology their citizens use.

Without a change, we may also miss the promise of these new technologies. While headline-grabbing warnings of the existential risks of AI have somewhat faded, democracies have for the most part retreated into their comfortable roles as regulators and rule-makers. The potential consequences are twofold: on the one hand, the impetus behind the development of blueprints for state-backed AI may be left to autocratic or authoritarian states, for which the potential of AI as a route to expanded power is irresistible. On the other hand, liberal democracies may fail to demand normative technology that meets the standards and needs of their countries, and may fail to take advantage of the power of AI to buttress liberal values and cultural norms or to transform public services. Without committed action to ensure its responsible development, AI may simply amplify the worst excesses of digital media and of state-led, technology-backed repression.

Competition, conflict and cooperation around the design, deployment and governance of emerging technology will remain central influences in global affairs. AI technology – on the battlefield and on the trading floor, in hospitals, newsrooms and classrooms – presents challenges and opportunities for states looking to advance their position in the world, or to respond to the concerns of citizens who expect their governments to protect and provide. Rising to this challenge – through clarity of mission and purpose, multi-stakeholder dialogue, and investment and innovation in governance – will ensure this latest technology is a force for global good. My hope is that this collection will drive that agenda forward.

5 International Telecommunication Union (2022), 'Facts and Figures 2022: Latest on global connectivity amid economic downturn', 30 November 2022, https://www.itu.int/hub/2022/11/facts-and-figures-2022-global-connectivity-statistics.

02
A ‘CERN for AI’ – what might an international AI research organization address?

The CERN nuclear physics laboratory was founded in 1954 to pool expertise and resources for fundamental research, in service of the common good, on a scale beyond the means of any single country. Might a similar model work for AI governance today, enabling its huge challenges to be tackled in an environment of depoliticized cooperation?

Elliot Jones

A moment for governance

The question of how to ensure artificial intelligence (AI) systems operate safely, ethically and legally is a challenging one. Risks and harms can originate and proliferate at different stages of an AI system’s life cycle. If poorly designed or hastily deployed, AI systems can cause a range of harms to people and society.[6] Job displacement, exacerbation of societal inequalities, and amplification of toxic content or racial stereotypes are all potential outcomes. Risks can also arise from misuse or catastrophic accidents, from AI systems malfunctioning and causing injury, and from poor supply-chain practices.

Recent advances in generative AI systems and ‘foundation models’ (AI models capable of a wide range of tasks) have exacerbated concerns about these risks.[7] Powerful AI capabilities such as text or image generation are more readily accessible to everyday users. Among policymakers, civil society organizations and industry practitioners, advances in AI have created a sense of urgency about the need to govern these new technologies effectively.[8]

6 Davies, M. and Birtwistle, M. (2023), ‘Seizing the ‘AI moment’: making a success of the AI Safety Summit’, Ada Lovelace Institute, 7 September 2023, https://www.adalovelaceinstitute.org/blog/ai-safety-summit.
7 Jones, E. (2023), ‘Explainer: What is a foundation model?’, Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer.
8 Altman, S., Brockman, G. and Sutskever, I. (2023), ‘Governance of superintelligence’, OpenAI, 22 May 2023, https://openai.com/blog/governance-of-superintelligence.

Local and national governments around the world are grappling in different ways with the challenge of governing AI. In the US, several cities have passed laws prohibiting or restricting the use of facial recognition systems.[9] At a regional level, the European Union has just finalized its AI Act, a comprehensive product safety law that will affect many AI systems.[10] These distinct measures address a common challenge: how to govern a suite of technologies that affect different sectors, involve complex supply chains, operate across borders and can raise a wide variety of risks. Leading AI labs like Anthropic, Google DeepMind, Microsoft Research and OpenAI have joined calls for a coordinated international effort to improve safety.

A 'CERN for AI'

Efforts such as the UK's AI Safety Summit, and the follow-up AI Seoul Summit, have aimed to start an international discussion around AI safety and have led to the creation of national AI safety institutes in the UK, the US, Japan and Canada.[11] There is a shared ambition to create an international AI safety network, and these institutes have begun to sign bilateral cooperation agreements. However, there is not currently a single, coordinated global institution seeking to promote or research safer AI. Recent research has explored how different existing models for international governance might be applied to AI, including whether institutions such as the Intergovernmental Panel on Climate Change (IPCC) or the International Atomic Energy Agency (IAEA) have features that could be borrowed or adapted for use in this emerging field.[12]

One increasingly prominent proposal, circulating among some academics and policymakers, advocates the creation of an international coalition for AI research inspired by the scale and collaborative spirit of CERN.

One increasingly prominent proposal, circulating among some academics and policymakers, advocates the creation of an international coalition for AI research inspired by the scale and collaborative spirit of CERN, the European Organization

---

9 Dave, P. (2022), 'Focus: U.S. cities are backing off banning facial recognition as crime rises', Reuters, 12 May 2022, https://www.reuters.com/world/us/us-cities-are-backing-off-banning-facial-recognition-crime-rises-2022-05-12.
10 Council of the EU (2024), 'Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI', press release, 21 May 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai; Edwards, L. (2022), *Regulating AI in Europe: four problems and four solutions*, Expert opinion, Ada Lovelace Institute, March 2022, https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf.
11 AI safety is a broad term with multiple complex and contested interpretations. Some argue that it encapsulates technical research to ensure AI systems are robust, unbiased, transparent and aligned with human values. Others contend that it must incorporate ethical, legal and social considerations.
12 Maas, M. and Villalobos, J. (2023), *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, AI Foundations Report 1, 23 September 2023, https://doi.org/10.2139/ssrn.4579773.

for Nuclear Research.[13] After the devastation of the Second World War, there was an effort to rebuild European science. Leading scientists, including the Danish nuclear physicist Niels Bohr, lobbied European governments to establish an international laboratory devoted to particle physics. In 1954, CERN was established under the umbrella of UNESCO.[14] CERN is publicly funded by 23 member states (22 European states plus Israel).

CERN is both an international institution and a laboratory. It is famous for discoveries like the Higgs boson and for being the birthplace of the World Wide Web. Its primary function has been to provide the powerful and prohibitively expensive infrastructure and hardware – most notably the Large Hadron Collider – needed to conduct particle physics research. CERN has historically focused on fundamental science rather than on developing technical standards or benchmarks. Its significant resources have enabled it to fund and broker multinational research collaborations. In addition, CERN has been an explicit source of inspiration in the institutional design of organizations in molecular biology and astronomy.[15] For these reasons, some people have argued that a similar institution could tackle the complex challenges of AI safety.

CERN's broader legacy has been in enabling nations to work together on expanding scientific knowledge. Constructing its particle accelerators and detectors required member states to pool expertise and funding. This allowed them to achieve a scale and depth of scientific work that no single country could have reached alone. CERN represents the post-war ideal of science beyond borders in the service of discovery and peace.[16]

If a CERN-like organization for AI were to exist, with the function of coordinating international AI safety research, it would require several features. First, as with the actual CERN's particle accelerators and supercomputers, a CERN-like body for AI would need its own technical infrastructure to support computational

---

13 Kaspersen, A. (2021), 'Time for an Honest Scientific Discourse on AI & Deep Learning, with Gary Marcus', Carnegie Council for Ethics in International Affairs, 3 November 2021, https://www.carnegiecouncil.org/media/series/aiei/20211103-honest-scientific-discourse-ai-deep-learning-gary-marcus; Kelly, E. (2021), 'Call for a 'CERN for AI' as Parliament hears warnings on risk of killing the sector with over-regulation', Science Business, 25 March 2021, https://sciencebusiness.net/news/call-cern-ai-parliament-hears-warnings-risk-killing-sector-over-regulation; Coyle, D. (2023), 'Preempting a Generative AI Monopoly', Project Syndicate, 2 February 2023, https://www.project-syndicate.org/commentary/preventing-tech-giants-from-monopolizing-artificial-intelligence-chatbots-by-diane-coyle-2023-02; Phillips, J. (2023), 'Securing Liberal Democratic Control of AGI through UK Leadership', James W. Phillips' Newsletter, 14 March 2023, https://jameswphillips.substack.com/p/securing-liberal-democratic-control; 'Securing Our Digital Future: A CERN for Open Source large-scale AI Research and its Safety' (2023), online petition submitted via openPetition on 1 June 2023, https://www.openpetition.eu/petition/online/securing-our-digital-future-a-cern-for-open-source-large-scale-ai-research-and-its-safety#petition-main; *The Economist* (2023), 'How to worry wisely about artificial intelligence', 20 April 2023, https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence; and Scholl, G. (2022), 'We need a CERN for AI in Europe', *Humboldt Kosmos* magazine, Alexander von Humboldt Stiftung, interview with Professor Holger Hoos, 1 August 2022, https://www.humboldt-foundation.de/en/explore/magazine-humboldt-kosmos/by-courtesy-of-how-artifcial-intelligence-is-changing-our-lives/we-need-a-cern-for-ai-in-europe.

14 Wanless, A. and Shapiro, J. N. (2022), *A CERN Model for Studying the Information Environment*, Carnegie Endowment for International Peace, November 2022, https://carnegieendowment.org/2022/11/17/cern-model-for-studying-information-environment-pub-88408.

15 Specifically, the European Molecular Biology Lab and the European Southern Observatory have both drawn explicit inspiration from CERN. See Sutton, C. (2014), 'Fifty years of EMBO', CERN, 17 July 2014, https://home.cern/news/news/cern/fifty-years-embo; and Sutton, C. (2012), 'ESO and CERN: a tale of two organizations', *CERN Courier*, 27 September 2012, https://cerncourier.com/a/eso-and-cern-a-tale-of-two-organizations.

16 Heuer, R. (2014), 'A celebration of science for peace', CERN, 20 February 2014, https://home.cern/news/opinion/cern/celebration-science-peace.

research. This could include physical infrastructure such as data centres, high-performance computing resources, networking systems, and laboratory facilities tailored to AI work.

Social and organizational infrastructure would be needed to provide operational support, including to manage relationships with commercial labs and nation states, make platforms available for open and innovative research communication, and secure sustainable funding for international research networks and collaborations. A CERN-like body might also foster more interdisciplinary and international research collaboration on AI risks, enabling greater involvement of researchers from countries that are otherwise lacking in computational resources.

If this new organization were to study the safety of frontier AI models, it would also require privileged, structured access to state-of-the-art AI models from industry labs, and access to underlying 'training' datasets – used to train AI systems in different capabilities – and other critical materials relating to each model's design and operation. Leading labs, including OpenAI, Google DeepMind and Anthropic, have already made voluntary commitments to open their models to select researchers for the purposes of safety and independent evaluations, though it remains unclear how meaningful these commitments will be unless underpinned by hard regulatory requirements.[17]

A CERN for AI might be in a unique position to broker access to cutting-edge AI systems, allowing researchers to test and compare the safety, biases and robustness of models beyond what any single lab could achieve independently.

However, a CERN for AI might be in a unique position to broker access to cutting-edge AI systems, allowing researchers to test and compare the safety, biases and robustness of models beyond what any single lab could achieve independently.[18] The convention that underpins CERN grants it status as an intergovernmental organization, with the privileges and immunities that come with that, and provides for direct contributions from governments; all this insulates it from political pressures in a way not possible for even national labs.[19]

---

17 Clarke, L. (2023), 'OpenAI, DeepMind will open up models to UK government', *Politico*, 12 June 2023, https://www.politico.eu/article/openai-deepmind-will-open-up-models-to-uk-government; The White House (2023), 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI', 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; Criddle, C., Gross, A. and Murgia, M. (2024), 'World's biggest AI tech companies push UK over safety tests', *Financial Times*, 7 February 2024, https://www.ft.com/content/105ef217-9cb2-4bd2-b843-823f79256a0e.
18 Ho, L. et al. (2023), 'International Institutions for Advanced AI', arXiv:2307.04699, arXiv, last revised 11 July 2023, http://arxiv.org/abs/2307.04699.
19 Convention for the Establishment of a European Organization for Nuclear Research, 1 July 1953, https://legal-service.web.cern.ch/system/files/downloads/CONVENTION.pdf.

Strengths of a CERN for AI

Proponents of a CERN-like body for AI have called for its creation as a way to build safer AI systems, enable more international coordination in AI development, and reduce dependencies on private industry labs for the development of safe and ethical AI systems. Rather than creating its own AI systems, some argue, a CERN-like institution could focus specifically on research into AI safety.

Some advocates, such as computer scientist Gary Marcus, also argue that the CERN model could help advance AI safety research beyond the capacity of any one firm or nation. The new institution could bring together top talent under a mission grounded in principles of scientific openness, adherence to a pluralist view of human values (such as the collective goals of the UN's 2030 Agenda for Sustainable Development), and responsible innovation.[20] Similar sentiments have been repeated by other prominent actors in the AI governance ecosystem, including Ian Hogarth, chair of the UK's AI Safety Institute,[21] who argues that an international research institution offers a way to ensure safer AI research in a controlled and centralized environment without being driven by profit motive.[22]

Proponents of a CERN-like model also argue that such an institution could provide vital global public goods for AI safety, which profit-driven private companies might otherwise undersupply. Such goods could include: benchmarks to evaluate model robustness; auditing tools to increase accountability; and datasets to assess harmful biases. Providing all of this would require sector-wide collaboration between governments and AI companies. Crucially, this work could include research into an expanded definition and operationalization of 'AI safety' that would cover the full scale of harms that AI systems can cause; such research could be informed by a deliberative process involving a representative sample of humanity, not just commercial labs or academics.

Some proponents of a CERN for AI believe that it may also reduce dependency on private labs for AI safety research, and attract top researchers interested in pursuing projects of greater public benefit rather than those with purely commercial potential. Professor Holger Hoos has described a potential CERN for AI as a 'beacon' to 'attract talent from all over the world'.[23] This could create an alternative hub of expertise outside the private sector.

The existence of a different type of AI institution could provide academics and students with an alternative career option to joining big tech firms. It could also help address asymmetries in political power between industry and academic labs.[24] Currently, significant power in AI development accrues disproportionately to a handful of private labs. A publicly funded international research organization

---

20 Kaspersen (2021), 'Time for an Honest Scientific Discourse on AI & Deep Learning, with Gary Marcus'.
21 AI Safety Institute and Department for Science, Innovation & Technology (2024), 'Introducing the AI Safety Institute', policy paper, updated 17 January 2024, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.
22 Hogarth, I. (2023), '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.
23 Scholl (2022), 'We need a CERN for AI in Europe'. The UK's AI Safety Institute, for example, has already been able to attract senior staff from OpenAI and Google DeepMind.
24 Ho et al. (2023), 'International Institutions for Advanced AI'; Clark, J. (2023), 'AI Safety and Corporate Power – remarks given at the United Nations Security Council', Import AI, 18 July 2023, https://jack-clark.net/2023/07/18/ai-safety-and-corporate-power-remarks-given-at-the-united-states-security-council.

conducting safety research might be more resilient than private sector labs to economic pressures, and better able to avoid the risk of profit-seeking motives overriding meaningful research into AI safety measures.

Hurdles faced by a CERN for AI

Long timelines and cost overruns often plague ambitious big science collaborations.[25] Physics breakthroughs have required enormous hardware investments over years. For example, to build CERN's Large Hadron Collider, over 10,000 scientists and engineers from hundreds of universities and labs contributed to its design and construction over a decade.

But while current computer clusters for AI research have yet to require such large workforces, constructing data centres and network infrastructure at scale for a new institute will still take time, investment, and reliable access to currently undersupplied specialized chips for AI development. That said, the modular nature of graphics processing units (GPUs) and servers could allow for much faster scaling up of AI infrastructure than has been feasible in previous science megaprojects.

Challenges in AI safety also differ from those of particle physics, so addressing them may require more dynamic, distributed initiatives. For example, CERN itself primarily focuses on pure science. However, AI safety is as much a question of values, ethics and societal impacts as it is a matter of AI systems' technical capabilities. Focusing on purely technical evaluations of an AI model's performance can only reveal so much information about its use and potential outcomes. It would thus be essential to ensure that critical perspectives on the impacts and implications of AI are incorporated from the outset into any new institution's culture and mission. But even this may not be enough: some risks of AI will only become apparent when a particular application is deployed, and may prove challenging for a CERN-like body to address.

In other words, it is possible that a CERN for AI could address only a subset of the risks that AI systems pose. This is due to the vast range of challenges that AI systems can present to actors in multiple domains. Focusing only on technical, model-level fixes such as better learning from human feedback, for instance, could prove a distraction from other essential governance efforts, such as regulation, accountability and public engagement, all of which are also necessary for identifying and mitigating risks from AI systems. Care would need to be taken to involve diverse stakeholders, and to balance capabilities against controls. Inflated expectations for AI governance via a CERN-like model could backfire if they are not realistic about such an organization's inherent limitations.

Another hurdle could be the issue of information asymmetry between the private sector and any new institution. Given its likely focus on safer systems and providing public goods, as discussed above, rather than purely pushing forward AI capabilities, the new institute would be unlikely to control the most capable AI systems itself. It would therefore be dependent on information-sharing from commercial labs

25 di Castri, G. (2021), 'Planning, scheduling and controlling long term projects', *Academia Letters*, Article 1284, June 2021, https://doi.org/10.20935/AL1284.

to understand those systems (which will depend on proprietary data, model design and engineering insights, which commercial labs will want to keep to themselves), absent any information-sharing obligations placed upon them (e.g. as is mandatory to meet safety requirements in the aerospace industry). Even if internal developments in commercial AI labs (such as safety concerns) are published openly, there will be a delay between discovery and those findings being shared more widely and acted on. Being 'behind the curve' in terms of understanding and having access to the most capable systems may also make working in public bodies less attractive for some. To mitigate this risk, an incentive structure would need to be established that can compete with private industry to attract and retain researchers.

There are also worries that creating a CERN for AI may result in safety researchers working in less close proximity to leading commercial AI labs, thus reducing the ability of such researchers to monitor risks on the ground. It may be that the best safety research is conducted alongside cutting-edge AI research in the private sector, as this could enable a deeper understanding of the systems and processes of the labs involved.[26]

There are worries that creating a CERN for AI may result in safety researchers working in less close proximity to leading commercial AI labs, thus reducing the ability of such researchers to monitor risks on the ground.

These issues also raise the concern that a new CERN for AI could be influenced or captured by big tech firms. To date, the research carried out by such firms has far outpaced public sector capabilities, with the result that major tech companies currently hold disproportionate power by virtue of their resources, expertise and leverage. Preventing narrowly commercial interests from dominating a CERN for AI would require vigilant governance.

That said, the governance structure of CERN could provide a template for its AI-focused equivalent: CERN's multinational membership and interdisciplinary focus insulate it from capture by special interests, and provide a diversity of input to counter corporate influence. CERN is run by a council of its member states, with two delegates each (one representing government, the other national scientific interests); each member state has a single vote, and the council operates on a simple majority vote for decision-making.[27] This also ensures no single member state can abuse its position within CERN – and provides a measure of protection against risks associated with the actions of individual states, as seen in the council's suspension of Russia's scientific observer status in March 2022 after Russia's full-scale invasion of Ukraine.[28]

---

26 Ho et al. (2023), 'International Institutions for Advanced AI'.
27 CERN (undated), 'Our Governance', https://home.cern/about/who-we-are/our-governance (accessed 22 Feb. 2024).
28 CERN (2022), 'CERN response to the aggression against Ukraine', 8 March 2022, https://council.web.cern.ch/sites/default/files/c-e-3626_Resolution_re_Russia%20.pdf.

Researchers have also raised concerns that giving a centralized institution access to the advanced AI models of leading labs might compromise the security of those labs and models.[29] For example, effective access to design evaluations and benchmarks may require the ability to copy a given model, which could undermine the commercial interests of those labs and enable diffusion of those models before adequate testing. This may be less of an issue for mechanistic interpretability and similar research, which may not require access to the latest models.

Lastly, a CERN for AI would have to grapple with rising geopolitical tensions. It is arguably harder today to start an international governance body than it was in the era immediately after the Second World War. Most leading AI labs are based in the US and China, two countries that are arguably engaged in a 'new cold war' that is fuelling a technological arms race between them.[30]

A path forward

A CERN-like institution would not be a replacement for comprehensive national and local regulation and governance frameworks, which would need to address broader challenges such as harmful misuses of AI systems. What is interesting about the proposal, though, is the potential that a new international body could complement the creation of other international governance organizations and instruments, including standards-setting bodies, certification bodies, treaties and domestic legal frameworks.[31]

There is no perfect analogue for AI when it comes to governance, and as future AI safety summits approach, policymakers should evaluate proposals for new international institutions and consider what these can accomplish, building on the efforts of the already established AI safety institutes. A CERN for AI undoubtedly represents one credible possible model to advance targeted elements of AI safety research and provide public alternatives to private sector dominance. With ample resources and global collaboration, it could make valuable technical contributions.

However, we cannot and should not expect one governance model to address the full span of risks and harms from AI systems. It may be that institutions such as the International Civil Aviation Organization, IAEA or IPCC provide better models for solutions to international AI governance. We cannot overlook the risk that a CERN for AI may turn out to be too expensive, too cumbersome, or simply unnecessary. More research is needed to flesh out what the realistic objectives of this kind of institution might be, how it might work, and what kinds of challenges it will be best placed to solve. The path forward rests on collective insight, courage and care in steering AI's immense potential towards the common good.

---

29 Ho et al. (2023), 'International Institutions for Advanced AI'.
30 AI Now Institute (2023), 'Tracking the US and China AI Arms Race', 11 April 2023, https://ainowinstitute.org/publication/tracking-the-us-and-china-ai-arms-race.
31 Trager, R. et al. (2023), *International Governance of Civilian AI: A Jurisdictional Certification Approach*, white paper, August 2023, http://dx.doi.org/10.2139/ssrn.4579899.

03
Regulating AI and digital technologies – what the new Council of Europe convention can contribute

Future AI regulation needs to be global in reach yet agile enough to allow each jurisdiction to tailor laws to local circumstances. The Council of Europe's new AI treaty offers a binding framework for ensuring AI regulation upholds existing standards on human rights, democracy and the rule of law – not just in Europe, but in all countries that share the same values.

Thomas Schneider

Editor's note: The author of this essay is the chair of the Council of Europe's Committee on Artificial Intelligence, which negotiated the AI framework convention.

A generation-defining technology and its challenges

Artificial intelligence (AI) is not in itself a new phenomenon. AI is already at the core of most of our everyday digital tools, including social media platforms, anti-virus software, virtual assistants and navigation software. But the rapid rise of new 'generative AI' models – which can produce various types of content, including text, imagery, audio and video – has captured the headlines, leading to alarmed calls from some quarters for caution and even for bans on AI's use. Suddenly, AI isn't just running in the background: it is a disruptive power in need of attention.

New technologies often transform societies and economies, and may demand new governance models. The industrial revolutions of the 18th to 20th centuries offer parallels to what we are witnessing today. Then, too, the reaction to new technologies was sudden and mixed, ranging from euphoria to panic. In 1832, textile home workers in the Swiss region of Zürich set fire to a mechanical weaving factory out of fear of losing their jobs.[32] Wilhelm II, the last German kaiser,

32 Meyer, B. (2019), 'La révolte contre les machines à Uster' [The revolt against machines in Uster], blog, Swiss National Museum, 5 August 2019, https://blog.nationalmuseum.ch/fr/2019/08/revolte-contre-les-machines-a-uster.

is quoted as saying 'the car has no future, I believe in the horse',[33] while people chased cars in the street because they loved the smell of the exhaust fumes and the oil.[34]

Data has often been described as the 'new oil' of the digital revolution.[35] To extend the metaphor, perhaps AI systems are its 'new engines': machines that process data to power applications, with the promise (or threat) of automating or replacing repetitive or laborious cognitive work. This includes highly skilled work, from writing and translation to marketing and decision-making in specialized fields. The implications are clear: like generation-defining technologies of the past, AI-driven tools will drastically change societies and economies, leading to the elimination of professions and the emergence of new ones. AI tools will lead to shifts in the balance of economic and political power. They will challenge existing orders, both locally and globally. And as with any technological revolution, they will produce not only winners but also losers.

The long-term risks of AI development are still uncertain. But 20 years of digital technology, data capture and machine-processing all point to changes in almost all industries. Traditional services providers and products will be squeezed or forced out of the market by newer, more efficient ones. Automated decision-making can be mysterious, and if decisions are no longer comprehensible or predictable, challenges around the rule of law, liability and autonomy are likely to emerge. First movers in AI-driven fields may establish dominant market positions through economies of scale, building on the emergent data monopolies of the past decade. Should data-rich and resourceful private companies continue to lead the way, there is a risk that society will become more dependent on powerful tech giants for stewardship of education, social services and healthcare provision, or for management of complex transport systems and energy flows.

Regulating AI, regulating the uses of AI

Whether addressing short-term risk or long-term uncertainty, well-conceived AI regulation is a necessity. The question now is what effective AI regulation might look like. And for that, we might draw lessons from one of the last great technological upheavals: the development of the internal combustion engine.

Today, no single 'engine law' regulates all aspects and impacts of engines. Rather, and over time, we have created a sophisticated system of technical, legal and social norms that regulate the use of engines, depending on context. The focus of regulation is mostly not on the engines themselves, but on the machinery they power and the risks associated with its uses. We have different regulations for the people who operate machinery. We also have rules for the fuels and infrastructure involved.

---

33 Adler, M. (2019), 'Das Jahrhundert der fossilen Mobilität geht zu Ende [The century of fossil mobility comes to an end], *Böll.Thema* 19-3, July 2019, Heinrich-Böll-Stiftung, https://www.boell.de/de/2019/07/02/das-jahrhundert-der-fossilen-mobilitaet-geht-zu-ende.
34 Hänggi, M. (2008), 'Was wäre, wenn sich das Elektroauto durchgesetzt hätte' [What would have happened if the electric car had been pushed through], *Neue Zürcher Zeitung*, 1 August 2008, https://www.nzz.ch/folio/was-ware-wenn-sich-das-elektroauto-durchgesetzt-hatte-ld.1619986.
35 Arthur, C. (2013), 'Tech giants may be huge, but nothing matches big data', *Guardian*, 23 August 2013, https://www.theguardian.com/technology/2013/aug/23/tech-giants-data.

These multiple sets of rules vary according to the domain of application, and they can also differ from country to country. They necessarily take into account different levels of cultural tolerance of risks, different levels of aversity to regulation, different approaches to dealing with risk, and different cultural approaches to the role of the state and individuals in this process. In addition, the extent to which rules are harmonized between jurisdictions depends significantly on the area of application: for example, *local* road traffic regulation varies far more widely than *international* air traffic regulation.

However, this analogy is far from perfect. AI systems have many properties quite unlike those of physical engines. They are digital tools that can proliferate quickly, can be copied at will, and can be transported more or less instantaneously across national borders. Above all, they can evolve and learn. Their functioning is more abstract than that of engines, and at the same time more complex. Sometimes not even the creators of AI systems understand what the systems do or how they produce their results. Moreover, the same AI systems can be used in very different contexts and for different purposes.

Any set of rules for AI that seeks to do justice to the nature of AI, and to be appropriate to the risks, must be just as dynamic and agile as the technology itself.

Any set of rules for AI that seeks to do justice to the nature of AI, and to be appropriate to the risks, must therefore be just as dynamic and agile as the technology itself. AI technologies pose global issues across states and regions, and therefore require a concerted response. But a 'concerted' or harmonized response is not the same as a uniform one: we may need to develop a system of technical, legal and cultural norms for AI applications that is *at least* as differentiated as those for engines. These norms will need to be based not solely on the technical features and capabilities of each system, but also on the risks associated with its application in any specific context.

The need for a common framework

When discussing the need for 'new' rules of the game for AI, we must not forget that existing national and international norms – including those protecting fundamental rights, human dignity and democracy – are applicable to new technologies.

For a number of years, international organizations such as the OECD, the Council of Europe,[36] UNESCO and the International Telecommunication Union (ITU) have worked on AI to understand its challenges and identify regulatory gaps, and have developed various soft law instruments accordingly. For instance, since 2018 the

36 The Council of Europe is an intergovernmental organization with 46 member states (including the 27 members of the EU). It was founded in 1949 with a mandate to promote human rights, democracy and the rule of law.

Council of Europe has developed soft law instruments on, *inter alia*, the use of AI in the judicial system[37] and the human rights impacts of algorithmic systems.[38] Since 2021, the EU has also worked on an AI Act. Designed to regulate AI in the internal market of the EU while respecting fundamental human rights and democracy, the act was approved by the Council of the EU in May 2024.[39] It contains, among other protections, special safeguards for some general-purpose ('horizontal') systems capable of being adapted to many uses, and for AI tools and applications deemed high-risk.[40]

While the AI Act is a milestone in its own right within the EU, equally significant has been a parallel push by the Council of Europe to establish the building blocks of a global regulatory regime for AI. From 2019 to 2021, the Council's Ad Hoc Committee on Artificial Intelligence (CAHAI)[41] examined the feasibility and potential elements of a legal framework covering the development, design and application of AI. Drawing on multi-stakeholder consultations, this work was informed by the Council's own standards on human rights, democracy and the rule of law as these pertain to AI, as well as by equivalent standards elsewhere. As a consequence of the CAHAI's findings, in June 2022 the Committee of Ministers of the Council of Europe mandated the Committee on Artificial Intelligence (CAI) – a new committee superseding the CAHAI – to negotiate a binding international agreement on the development, design and use of AI. The terms of this mandate required the framework to be based on the Council's existing norms on human rights, democracy and the rule of law while also being conducive to innovation.

A common framework with a global reach

From the beginning, the ambition of the Council and its member states had been to develop not just a legal framework for Europe, but the first legally binding international AI treaty of global reach. The idea was that such a treaty, though European in origin, would be open to any countries that uphold the principles of human rights, democracy and the rule of law.

This ability of a European treaty to shape global AI governance is supported by precedent in other domains. Council of Europe frameworks have a history of success: the Convention on Cybercrime (2001), 'Convention 108' on data protection (1981) and its revised version, 'Convention 108+' (2018), provide exemplary global

---

37 Council of Europe (undated), 'CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment', https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.
38 Council of Europe (2020), 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems', 8 April 2020, https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.
39 Council of the EU (2024), 'Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI', press release, 21 May 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai; European Parliament (2024), 'Artificial Intelligence Act: MEPs adopt landmark law', press release, 13 March 2024, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.
40 Ibid. See also Gibney, E. (2024), 'What the EU's tough AI law means for research and ChatGPT', *Nature*, 16 February 2024, https://www.nature.com/articles/d41586-024-00497-8.
41 Council of Europe (undated), 'CAHAI – Ad hoc Committee on Artificial Intelligence', https://www.coe.int/en/web/artificial-intelligence/cahai.

vehicles for cooperation among around 100 states. Such instruments are binding intergovernmental agreements which democratic states around the world can sign up to.

Input from diverse stakeholders shaped the drafting of the AI framework convention from the outset. A number of non-European states participated in its early development, while others joined later during the negotiations. (By the time the CAI had agreed a draft treaty on 14 March 2024,[42] after 19 months of intense negotiations, the list of official non-European participants consisted of Argentina, Australia, Canada, Costa Rica, Israel, Japan, Mexico, Peru, the United States and Uruguay.) The CAI also included observers from civil society, academia, the business sector and the technical community. Reflecting the rationale that individual states would need the approval of their parliaments to ratify the convention, a subgroup of potential future state parties was created and tasked with drafting the articles of the convention. Drafts were presented and explained in plenary sessions to all stakeholders, who were able to submit their own written and oral comments and propose text changes before and after every drafting group meeting. In this way, wide-ranging input and feedback on the text of the draft convention were ensured from all stakeholders until the very end of the negotiations.

Contrary to some expectations, the negotiating parties intended neither to create substantive new human rights nor to undermine the scope and content of existing applicable protections. Instead, they agreed a set of legally binding obligations and principles under which *each party's* existing applicable obligations in respect of human rights, democracy and the rule of law would be applied in the context of the new challenges raised by AI. Agreement by all parties on the need for a *graduated and differentiated* approach was important for ensuring that any future regulation and related measures would address, and be proportionate to, context-specific risks and impacts.

It was also clear that a framework convention designed to set the tone for AI governance in many jurisdictions over the coming decades could never anticipate, and was not intended to regulate, all aspects of AI in detail. Rather, it was (and is) meant to be supplemented – as in the previously mentioned illustrative example of 'engines' – by further technical, legal and sociocultural norms on aspects of AI used in specific contexts and countries. These norms will need to be developed and adapted continuously in each country or jurisdiction.

In addition to bridging gaps between legal systems, one of the biggest challenges during the negotiations for the AI convention had been to manage the expectations of some European states and civil society actors on regulation of the private sector in important *non-European* nations. Governments and civil society in Europe needed to realize that it is not possible simply to transfer the European system and its unique logic – based on the European Convention on Human Rights and the European Court of Human Rights in Strasbourg – to a global instrument. In order

42 Council of Europe (2024), 'Artificial Intelligence, Human Rights, Democracy and the Rule of Law Framework Convention: Statement by Secretary General Marija Pejčinović Burić on the occasion of the finalisation of the Convention', 15 March 2024, https://www.coe.int/en/web/portal/-/artificial-intelligence-human-rights-democracy-and-the-rule-of-law-framework-convention.

for the new AI convention to become an instrument of global reach, it needed to leave as much flexibility as possible for potential future parties to implement its principles while remaining compliant with their own national legal and regulatory frameworks. The more flexible the convention, in other words, the more countries would likely be able (and willing) to accede to it. Notwithstanding these factors, consensus on a set of core principles remained essential for brokering agreement and ensuring the alignment of signatories.

In order for the new AI convention to become an instrument of global reach, it needed to leave as much flexibility as possible for potential future parties to implement its principles while remaining compliant with their own national legal and regulatory frameworks.

Despite several critical moments during the negotiations, when it seemed impossible to bridge differences between the expectations of some European states/stakeholders and the realities in other countries, in the end the will and commitment on all sides to draw up an agreement with a global reach prevailed. The Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law was adopted in Strasbourg on 17 May 2024.[43] The convention obliges all future parties to address the risks from activities by both public and private actors within the lifecycle of AI, taking into account the respective roles and responsibilities of all stakeholders. It gives parties the flexibility to meet their obligations under the convention according to their own domestic legal and institutional frameworks. A periodic reporting mechanism will cover the measures taken by each signatory; this should both increase the accountability of states and help to ensure a dynamic approach to AI in the future. The convention's follow-up mechanism will also offer new opportunities for cooperation with states that have not yet ratified the treaty – this will further contribute to its potential global reach.

Beyond a common framework

Yet while establishing a common language and a binding commitment to shared values and fundamental principles is a necessary first step for regulating AI globally, it is not a sufficient one. New technologies demand agile and adaptive approaches to their governance. New AI applications are emerging every day. The boundaries between the state and the private sector, between the national and international, between different sectors, and even between science and business are becoming increasingly blurred. Beta versions and trial applications of AI are almost certain

43 Council of Europe (2024), 'Council of Europe adopts first international treaty on artificial intelligence', press release, 17 May 2024, https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence.

to have seismic effects on societies, and the speed of change may present difficulties for rigid and slow-moving decision-making bodies. Many of society's governance mechanisms have barely changed in decades or even centuries, and are reaching their limits in terms of keeping up with the evolution of digital technology.

Debate continues on the forms that AI governance and regulation might best take. Solutions could include the use of observatory models, risk mitigation, standards or watchdogs, among other options. However, this author's instincts and the evidence from other successes in technology governance suggest that best practice might include the following elements: interdisciplinary, multi-stakeholder processes; the establishment of sector- and application-specific regulatory priorities; and dynamic and agile legislative and executive processes that embrace the logic of the digital revolution rather than rejecting it. Digital governance must develop in smaller and faster steps, perhaps through regulatory 'updates' or 'releases' – similar to those seen in software – that react to technical developments immediately. We may even need to use AI systems themselves to develop regulatory frameworks that can cope with AI.

Whatever options are considered, given the fast-evolving and transnational nature of AI, collective governance requires shared international values and norms as well as binding commitments to respect and live up to them. The Council of Europe convention on AI provides a compelling route to get us there.

04
Community-based AI

AI need not inevitably be the domain of Big Tech. Smaller-scale, community-led work, dedicated to solving local problems and empowering marginalized groups, can have real impact. By embedding cultural and linguistic diversity into AI, community-based approaches could enable globally equitable outcomes and help counter the technological monoculture of big business.

Kathleen Siminyu

If we have learned only one thing from a decade and a half of social media, it might be that technology designed, built and operated in just one place but deployed worldwide should give us pause for thought. Values, assumptions and rules written into the technology we use matter enormously. Global technology frequently just means Western technology.

The development of technology enabled by artificial intelligence (AI) risks following a similar path. Of the many challenges this creates around ensuring AI is developed to the benefit of all, two in particular stand out. The first is the possibility that governance of so-called 'global' technology will map poorly to reality in geographies or cultures unfamiliar to its creators and operators. The second is that dominant, Western-developed technologies may squeeze out other technologies built by those very same under-represented geographies and cultures.

In both cases, this would not only hurt local communities in numerous ways (from entrenching cultural biases to limiting the creation of AI solutions relevant to local needs). It would also be a loss for AI development globally, limiting the pool of technical talent for AI work and inhibiting the diversity of perspectives and technical approaches needed to drive innovation.

'Global technology' and the linguistic dominance of English

ChatGPT, the OpenAI product that thrust generative language models into headlines and fuelled the widespread use of chatbots, is a good example of what could be considered a 'global technology'. ChatGPT is a large language model (LLM),

a type of AI that is created by trawling huge datasets of text with the aim of generating human-like responses (i.e. resembling examples in the text datasets) when prompted with questions or comments in a conversational manner.

The magic tricks that LLMs can perform in English are frequently astonishing. LLMs can generate sophisticated and (at least superficially) plausible text that is often indistinguishable from that written by a human. But LLMs are far more useful to English speakers than to anyone else. This is due to, on a surface level, issues in the web-crawled data for low-resource languages;[44] and fundamentally, systemic issues in society that are then reflected in the lack of availability of data for these languages, and in the poor quality of the data on the occasions when it is available.[45] Low-resource languages are languages for which insufficient data is available to enable development of robust natural language processing (NLP) capabilities in AI systems. One study found that for at least 15 such poorly represented languages, the data used to train LLMs was totally deficient.[46] In other words, for non-English communities, there is a higher likelihood that AI tools nominally developed for a given language might actually spit out gibberish that is 'like' the language in question; this is in addition to the AI having minimal factual knowledge of the local contexts in which the language is spoken.[47]

Other researchers evaluating the performance of LLMs developed for a global audience in relation to that of LLMs developed to serve subsets of African languages found that global AI tools are 'still not achieving the accuracy of low-resource and Africa-centric language models, [even] on simple tasks…'.[48] For example, an evaluation by Lelapa AI – a South Africa-based AI lab – of ChatGPT's performance in Zulu found that LLMs built by native-language teams and focused on a subset of languages performed significantly better at named entity recognition than LLMs developed with a global scope did. (Named entity recognition is a crucial step in information extraction, and is used to classify proper nouns in formerly unstructured text.) For machine translation, the gulf was even wider, with ChatGPT 3.5 scoring a round zero as its BLEU score.[49] The team concluded its study by stressing 'the huge value of context-specific AI work'.[50]

Colonial AI, or local AI?

In principle, universal, global AI, should such a thing be possible, would have none of these problems. Yet the above-mentioned issues with output quality, combined with prevailing economic systems that leave dominant populations as the owners of such tools, suggest a risk that the globalization of AI could facilitate what might

---

44 Kreutzer, J. et al. (2022), 'Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets', *Transactions of the Association for Computational Linguistics*, Volume 10, MIT Press, pp. 50–72, https://aclanthology.org/2022.tacl-1.4.
45 Nekoto, W. et al. (2020), 'Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages', *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–60, Association for Computational Linguistics, https://aclanthology.org/2020.findings-emnlp.195.
46 Kreutzer et al. (2022), 'Quality at a Glance'.
47 Deck, A. (2023), 'We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed.', *Rest of World*, 6 September 2023, https://restofworld.org/2023/chatgpt-problems-global-language-testing.
48 Abbott, J., Dossou, B. and Mbuya, R. (2023), 'Comparing Africa-centric Models to OpenAI's GPT3.5', Lelapa AI, 9 February 2023, https://lelapa.ai/comparing-africa-centric-models-to-openais-gpt3-5-2.
49 A 'BLEU' (Bilingual Evaluation Understudy) score is a means to compare how a text has been translated by an automated system compared to the original references created by human translators.
50 Abbott, Dossou and Mbuya (2023), 'Comparing Africa-centric Models to OpenAI's GPT3.5'.

be termed a form of 'colonial AI', with its insistence on English and little regard paid to local cultures and languages. The near-invisibility, at least until recently (see below), of non-Western and non-English-speaking AI stakeholders in much of the AI debate is also evident in the fact that international AI conferences are typically held in the Global North; attendance of African delegates is usually low, due to distance and high travel costs and registration fees.

There are also concerns that globalized AI could lead to or entrench exploitative economic dynamics, as suggested by reports on the human cost of preparing high-quality training data for building AI models. Data annotation (the practice of human coding and labelling of text, images or videos) is usually outsourced, and often poorly paid. This labour is essential for limiting bias, hate speech, violence and sexual abuse content generated by AI models. However, it has been reported that workers in this field often endure difficult working conditions and are exposed to toxic textual and visual content, which affects their mental health. Even though outsourced data annotation is, in a sense, the backbone of a highly lucrative industry, this is seldom reflected in the status, compensation and protections provided to those performing such roles.

If AI content continues to be developed in such ways, it will have significant negative ramifications for both its producers and its consumers. Failings in the governance of social media have led to violence in parts of the world where companies have not invested in appropriate oversight or care. AI companies need to avoid repeating this mistake: a risk assessment for a model in the UK or US, for example, should not mirror assessments for Bangladesh or Kenya. With AI tools potentially acting as news sources, personal assistants, recruiters or political advisers, it should be a critical priority to ensure each tool is safe and fit for use in a given culture or country.

There is encouraging evidence of a growing African AI community that is taking ownership of AI development and the issues around it, and building and shaping AI technologies that respond to local needs.

Fortunately, the story of AI development in the future is unlikely to be confined to Western companies and cultures. Quite the opposite. For example, there is encouraging evidence of a growing African AI community that is taking ownership of AI development and the issues around it, and building and shaping AI technologies that respond to local needs. Women in Machine Learning and Data Science, an organization that champions opportunities for women and gender minorities in these technical fields, has locally organized chapters in 13 African countries.[51] Data Science Africa,[52] Data Scientists Network (formerly Data Science Nigeria)[53] and the Deep Learning Indaba[54] are grassroots organizations championing capacity

51 Women in Machine Learning & Data Science, https://wimlds.org/chapters.
52 Data Science Africa, https://www.datascienceafrica.org.
53 Data Scientists Network, https://www.datasciencenigeria.org.
54 Deep Learning Indaba, https://deeplearningindaba.com/2023.

development in the African AI community. These groups organize events, summer schools, boot camps and conferences where members of the African AI community nurture the interest of budding young developers and support academic careers.

An increasing number of higher education opportunities in Africa are emerging in AI and machine learning: the African Institute for Mathematical Sciences (AIMS) runs a master's degree course in machine intelligence;[55] Google DeepMind provides scholarships for students at Stellenbosch University[56] in South Africa and Makerere University in Uganda;[57] and the African Centre for Technology Studies offers AI4D doctoral scholarships to candidates from 21 African countries.[58] Organizations like the Masakhane Research Foundation,[59] GhanaNLP,[60] EthioNLP[61] and HausaNLP[62] are also providing capacity-building, all working to increase NLP research on African languages and with various regional or linguistic focuses.

This flourishing AI ecosystem has had profound effects on the nature of technology in African countries. Language models that reflect local communities are being built. Problems that have previously received little attention from the teams and communities that traditionally develop AI in the West are now being placed front and centre. One example is treatment of *Leishmaniasis*, a neglected disease most common in Brazil, East Africa and India. Closely associated with poverty, the disease has historically received limited funding for discovery, development and delivery of new treatments. Moreover, the pre-existing treatment was costly, lengthy, painful and sometimes toxic.

In 2021, the 'Deep Learning Indaba Grand Challenge'[63] focused on this disease, bringing in local AI practitioners to lead model-building and assist with drug discovery and data analysis. Together, the participants started to tackle a disease that might have been ignored by the mainstream. Over 350 community volunteers participated in the challenge. This led to the selection of several promising drugs that went on to be evaluated by the Drugs for Neglected Diseases initiative – DNDi.[64] The winning solution from the Grand Challenge was published as a conference paper at the International Conference on Learning Representations (ICLR) in Vienna in 2021.[65]

Further evidence that the African AI community's international profile is increasing – and that practitioners may gain a greater say in the global development of AI – can be seen in the fact that in 2023 the ICLR was held in Africa for the first time. African

---

55 African Institute for Mathematical Sciences, https://aimsammi.org.
56 'DeepMind scholarships for postgraduate studies in machine learning at Stellenbosch University', https://mlai.sun.ac.za/dms.
57 'DeepMind scholarships at Makerere University', https://cs.mak.ac.ug/funding/scholarships/2023/deepmind.
58 African Centre for Technology Studies (2022), 'AI4D Scholarships Enhancing Doctoral Training in Artificial Intelligence (AI) and Machine Learning (ML) in Sub-Saharan Africa', https://www.acts-net.org/ai4d-background.
59 Masakhane Research Foundation, https://www.masakhane.io.
60 Ghana NLP, https://ghananlp.org.
61 Ethiopian NLP, https://www.ethionlp.com.
62 HausaNLP (undated), 'Hausa-NLP Open Community', https://github.com/hausanlp.
63 Deep Learning Indaba (2021), 'Indaba NDABA Grand Challenge: Curing Leishmaniasis', https://deeplearningindaba.com/grand-challenges/leishmaniasis.
64 Drugs for Neglected Diseases initiative, https://dndi.org.
65 Dassi, L. K., Kane. H. and Nkwate, E. (2021), 'Computationally accelerating protein-ligand docking for neglected tropical diseases: A case study on drug repurposing for leishmaniasis', conference paper, ICLR 2021, https://africa.ai4d.ai/wp-content/uploads/2021/05/ICLR_2021____Drug_Repurposing_Deep_Learning__Practical_ML_for_Developing_Countries.pdf.

attendance at the conference in Kigali, Rwanda grew by over 1,000 per cent. While these numbers are a testament to the increase in AI activity on the African continent, the geographic accessibility of the conference certainly played a role.

Open but vigilant

Building local capacity is the foundation of an anti-colonial technology movement, and it is delivering results. But protecting the interests of historically marginalized communities and their technology requires more than simply moving with the times. Steps can be taken to further strengthen the power and autonomy of small technology communities. Licensing the data created and curated by these communities is one example. It would be sadly ironic if such data were simply sucked up by a technology giant and then sold back to the very people who generated it.

The case of the Kaitiakitanga licence offers a positive example of what can be achieved. Te Hiku Media, a collectively owned charitable media organization, started gathering data for the Te Reo language, primarily spoken by the Māori of New Zealand but at risk of extinction following British colonial policies. Te Hiku Media noted the risk that, if such data was left unprotected, foreign technology companies might simply be able to develop products and sell these back to the Māori people. To address this risk, Te Hiku Media developed the Kaitiakitanga licence,[66] designed to ensure that access to the language dataset and any related resources aligns with the customs, protocols and values of the Māori people.

Risks and opportunities

Without addressing colonial power structures recreated in the design and deployment of technology, there is a risk that old inequities will find new life in the technological tools used by local communities, with consequences for such communities' power and identity. Displacement of local language and culture from our technology threatens communities' futures, cultural heritage and indigenous knowledge. The threat is a future in which local identities are erased by technological development. In short, there is a risk of dependency on tools and processes built by someone else, for someone else, for purposes that fail to deliver good outcomes for those excluded from their design and value chain.

In contrast, community AI can provide a route to digital self-determination. On local questions, community-based AI is likely to outperform global solutions and to be preferred by local populations; it simply needs to overcome the hurdle of getting the right tools into the hands of the people it is built for. In turn, community AI promises to bring value back to communities as the builders and the owners of these technologies.

66 Papa Reo API Kaitiakitanga License, https://papareo.io/kaitiakitanga.

05
Open source and the democratization of AI

Until recently, AI has been developed in the open. Now, risk aversion and commercial imperatives are reversing this trend. But use of systems built solely behind closed corporate doors would bring unwelcome centralization of control, and could result in unequal distribution of AI's benefits. Open-source AI must be allowed to thrive.

Alek Tarkowski

The meteoric rise of artificial intelligence (AI) in political consciousness has come alongside a major shift in how the technology is being built. Once a technology developed along open-source principles, AI is increasingly hidden away on grounds of safety, intellectual property rights or defence of trade secrets. This shift must be counteracted. From local councils to libraries, schools to universities, the power of AI-enabled technologies to transform our lives and the services we depend on is enormous.

Limiting AI's development to only the most powerful corporations would be a major setback in ensuring its benefits are felt as equitably as possible. Proportional regulation, protections for open-source data, and public sector skills and investment to secure a place for public AI alongside private AI are all necessary.

Open beginnings

Modern AI development is founded on the principle of openness. For decades, AI was primarily a research discipline, existing both in academia and industry. AI research teams in companies openly shared innovations. Even today, TensorFlow[67] and PyTorch,[68] the two critical machine-learning frameworks, built by Google and Facebook (currently Meta) respectively, remain shared as open-source code. Similarly, the Transformer architecture,[69] a novel and

67 TensorFlow, https://www.tensorflow.org.
68 PyTorch, https://pytorch.org.
69 Vaswani, A. et al. (2023), 'Attention Is All You Need', arXiv.org, last revised 2 August 2023, https://arxiv.org/abs/1706.03762.

widely used approach to deep learning, is an open-source innovation shared by Google Brain engineers. Such methods, combined with an open publishing culture embracing the use of preprint archives such as arXiv,[70] have been crucial in allowing researchers to share ideas.

As recently as 2017, open-source approaches still seemed to be in the ascendant. Nick Bostrom, one of the leading thinkers in providing an ideological underpinning for today's AI development, observed that 'leading AI developers operate with a high degree of openness'.[71] Analysing the strategic implications of openness – which he understood to mean the sharing of public-domain source code, scientific discoveries and AI platforms – Bostrom concluded that the short- and mid-term effects of openness would most probably be net positive.[72] That same year, OpenAI launched as a non-profit initiative. Openness was explicit in its brand identity.

Closing up

Seven years later, things look different. Today, OpenAI is one of several commercial giants offering closed and non-transparent AI systems in an increasingly concentrated market. A company manifesto from February 2023 states that 'we were wrong in our original thinking about openness' and frames the new approach as being to 'safely share access'.[73] In early 2024, the French AI startup Mistral followed the same trajectory. Although it was launched in mid-2023 as an open-source AI lab, the company decided not to release its latest model, Mistral Large, openly. Critics have questioned whether such shifts are as much about safety as they are about technology companies protecting their market value, but safety is the touchstone for many arguing against openness in AI research. A recent op-ed in the *Financial Times* compared machine learning with pathogen research,[74] a field premised on mitigating risk at any cost; the article was one of a number of voices highlighting the risks of working with AI in the open. A widely shared white paper written by DeepMind researchers offers a taxonomy of risks related to the operation of language models; these risks include discrimination, the spread of hate speech, misinformation, bias and exclusion.[75] The underlying argument here is that open systems lack control mechanisms to mitigate risk.

GPT-2, the first generative language model that found the limelight, was not immediately open-sourced. OpenAI argued that its decision to develop GPT-2 using a more closed approach was based on ethical considerations, and on the potential risk that the model would be used to create 'deceptive, biased, or abusive language at scale'.[76] With the deployment of the next generations of its model,

---

70 arXiv, https://arxiv.org.
71 Bostrom, N. (2017), 'Strategic Implications of Openness in AI Development', *Global Policy*, pp. 135–47, 9 February 2017, https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12403.
72 Ibid.
73 Altman, S. (2023), 'Planning for AGI and beyond', OpenAI blog, 24 February 2023, https://openai.com/blog/planning-for-agi-and-beyond.
74 Tett, G. (2023), 'The perils of open-source AI', *FT magazine*, 14 June 2023, https://www.ft.com/content/0cad55cd-7f07-4fd6-86b7-a2bbfacd214c.
75 Weidinger, L. et al. (2022), 'Taxonomy of Risks posed by Language Models', FAccT, June 2022, pp. 214–29, https://doi.org/10.1145/3531146.3533088.
76 OpenAI (2019), 'Better language models and their implications', OpenAI, 14 February 2019, https://openai.com/research/better-language-models.

GPT-3 and GPT-4, OpenAI has moved further still from openness and does not even provide basic documentation of these systems. Google is similarly not sharing its innovations, and gated API access is becoming the standard for AI services made available to the public. In 2023, Meta launched its Llama model (followed by Llama 2, later that year) using a hybrid approach. The code was openly released, but there were legal limitations on its reuse.

The driving forces behind limiting access to AI for reasons of safety – regardless of the motivation – are often the big AI industry players. Their requests to regulate and license AI development may be presented as solutions to AI risks, but critics argue that this limits market competition.[77] The most common narrative equates closed AI with responsible AI, and open models with AI risk. This is a line of reasoning repeatedly presented by industry, and picked up by the US government in its negotiation of voluntary commitments from AI companies. Europe has taken a different path with its new AI Act.[78] This regulation focuses on mitigating high-risk AI systems, and includes general-purpose AI models – deemed by many to be riskier because of the wide range of applications – in its scope. Yet carve-outs to obligations placed on AI developers have been included for open-source AI,[79] with policymakers recognizing the benefits of open development and deployment of AI.

Requests to regulate and license AI development may be presented as solutions to AI risks, but critics argue that this limits market competition.

Nonetheless, while the regulatory trend is still not definite, the dominant narratives supporting closed approaches are a cause for concern. Above all, they fail to account for a systemic risk that openness can mitigate: the risk of centralized control of powerful technologies and the monopolization of beneficial outcomes of AI systems. This is a risk that Bostrom noted in his paper.[80] A report from the UK Competition and Markets Authority, published in April 2024, outlines risks to competition on AI foundation models.[81] It is telling that the document comes from a market regulator, rather than from an AI policy institute. Such issues are often largely ignored in policy debates. For example, the DeepMind risk taxonomy

---

77 Kapoor, S. and Narayanan, A. (2023), 'Licensing is neither feasible nor effective for addressing AI risks', AI Snake Oil, 10 June 2023, https://www.aisnakeoil.com/p/licensing-is-neither-feasible-nor.
78 Council of the EU (2024), 'Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI', press release, 21 May 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai; European Commission (2021), 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts', COM/2021/206 final, 21 April 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206; European Parliament (2024), 'Artificial Intelligence Act: committees confirm landmark agreement', press release, 13 February 2024, https://www.europarl.europa.eu/news/en/press-room/20240212IPR17618/artificial-intelligence-act-committees-confirm-landmark-agreement.
79 Tarkowski, A. (2024), 'AI Act fails to set meaningful dataset transparency standards for open source AI', Open Future, 7 March 2024, https://openfuture.eu/blog/ai-act-fails-to-set-meaningful-dataset-transparency-standards-for-open-source-ai.
80 Bostrom (2017), 'Strategic Implications of Openness in AI Development'.
81 Competition and Markets Authority (2024), 'CMA outlines growing concerns in markets for AI Foundation Models', press release, 11 April 2024, https://www.gov.uk/government/news/cma-outlines-growing-concerns-in-markets-for-ai-foundation-models.

mentions the risk of concentration of power only indirectly.[82] Yet the concentration of power is, in fact, a fundamental AI risk that can only be mitigated by measures to decentralize and democratize access to, and use of, these technologies.

The resilience of open-source systems

The move towards closed models is far from a fait accompli. In July 2022, the Big Science consortium released BLOOM, a fully open, large language model comparable to GPT-3. BLOOM has open-source code, transparent training datasets, and a collaborative production model that involved over 1,000 researchers. In the same month, Stability.AI released Stable Diffusion, a fully open text-to-image model that could produce images similar to those generated by proprietary models such as Midjourney or Dall-E. While BLOOM and Stable Diffusion have attracted mainstream attention, many other open-source solutions have been developed in recent years. These include Pythia, a model built by the Eleuther.ai non-profit that allows researchers to better understand how AI models work, and StarCoder, a family of language models for computer code.

Together, these examples of open-source models signal the possibility of democratizing and decentralizing AI development. They demonstrate that a different trajectory is possible than that of centralization through proprietary solutions. Just as with browsers and operating systems in the past, open-source solutions have become viable alternatives and challenges to a potential AI oligopoly. Today, a robust field of open-source AI science is leading in areas such as training dataset creation, security research and model fine-tuning, and the models being built can be freely applied to non-commercial applications, large or small.

Decentralizing AI power should be a policy goal in itself, comparable to anti-trust efforts. Open-sourcing AI, as a decentralization method, would increase market competition. And while the creation of new AI models is prohibitively expensive, their further development and fine-tuning can often be conducted at much lower cost, offering a business model for market entrants and smaller, less resourced companies.

Open-source AI also has all the benefits associated with open research, by giving researchers broad, equal access to the technologies involved. Despite other mainstream narratives, the open-sourcing of models allows for greater scrutiny, and therefore helps solve issues such as bias, security or environmental concerns.

Open-source approaches can also help efforts to diversify AI technologies and make them available to people around the globe. Large AI firms have a history of treating languages from around the world as raw resources that can be extracted, exploited and enclosed in proprietary systems (see also Chapter 4, ‘Community-based AI’, and Chapter 6, ‘Resisting colonialism – why AI systems must embed the values of the historically oppressed’). The reverse trend is being championed by open-source developers. The collaboratively built BLOOM model works in 46 languages, and is based on justly sourced data. The open-source Polyglot-Ko is currently the best

82 Weidinger et al. (2022), ‘Taxonomy of Risks posed by Language Models’.

language model that works in Korean. Another example is Te Hiku Media (see also Chapter 4), a Māori organization that uses open-source technology to build sovereign AI solutions that both preserve and protect Māori language and tribal knowledge.

Today, debates about AI focus on the development of technologies. We are still in the early phases of their deployment, for example through chatbots like ChatGPT or Claude. Yet as AI solutions become more ubiquitous, we will either have the choice of a variety of solutions or a single corporate offer. This will be a choice faced by every small business, every non-profit organization and every school system. The market is already skewed. Today, any client of AI services most probably pays one of several providers, and indirectly pays for the services of an even smaller set of cloud companies that provide necessary computing power. In the field of AI services, market competition will also mean democratization.

What now?

Policymakers face a choice. As Frank Pasquale, a law professor and AI expert, has observed, one strategy could be to accept – or even promote – 'digital gigantism' and focus on regulating it.[83] This is expressed in calls for licensing AI developers or focusing on AI safety in close cooperation with commercial AI giants.

This strategy may be building momentum, but it is not without challenge. Around the world, governments and similar administrations are recognizing that it is necessary to prioritize both building on the strengths of open AI science and ensuring the freedom of non-commercial players to create and deploy AI systems for non-commercial needs. Governments from the US to Sweden, and bodies such as the EU, are receptive to the importance of open AI science. This is a major reason for optimism.

The introduction of exceptions for open-source AI systems within the scope of the AI Act[84] is a symbolic first. The key elements of these amendments include provisions supporting open-source AI development and rules for increased transparency and governance of training data. Although these measures have been lacking from the voluntary commitments secured by the US government from the seven largest AI companies, the bipartisan CREATE AI Act ('Creating Resources for Every American To Experiment with Artificial Intelligence Act of 2023') pushes for 'a shared national research infrastructure that provides AI researchers and students from diverse backgrounds with greater access to the complex resources, data, and tools needed to develop safe and trustworthy artificial intelligence'.[85]

---

83 Pasquale, F. (2018), 'Tech Platforms and the Knowledge Problem', *American Affairs*, Summer 2018, at 3, U of Maryland Legal Studies Research Paper No. 2018–19, 20 June 2018, https://ssrn.com/abstract=3197292.
84 Council of the European Union (2024), 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – Analysis of the final compromise text with a view to agreement', 5562/24, 26 January 2024, https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.
85 Congresswoman Anna G. Eshoo (2023), 'AI Caucus Leaders Introduce Bipartisan Bill to Expand Access to AI Research', press release, 28 July 2023, https://eshoo.house.gov/media/press-releases/ai-caucus-leaders-introduce-bipartisan-bill-expand-access-ai-research.

Striking a balance between commercial and non-commercial AI, open and closed AI, and safety and opportunity should be at the heart of AI policy. This means proportional regulation, strong data rights and public involvement. Three specific principles can be advocated:

Firstly, **policies should support open-source AI development** by making sure that regulation is proportional and does not unduly burden developers. In particular, proposals for licensing AI developers run the risk of concentrating development in the hands of major players. Self-governance practices developed in open-source projects – especially practices ensuring documentation and transparency – can serve as blueprints for regulation.

Secondly, **policies must acknowledge that AI development depends on a robust corpus of data**. While the spotlight for most of today's policy debates is on the governance of AI models (e.g. their alignment with human values, accountability and responsible use), governance of training data is also a fundamental aspect of AI policy. The legal status of training AI to perform certain tasks (such as generating text, for instance) using content and data taken from the open web is currently unclear and depends on the jurisdiction. European text- and data-mining exceptions allow such 'scraping' of internet sources; in the US, fair-use status is currently being tested in courts. On the other hand, the practice has understandably met with opposition from creators and rightsholders in the creative sector. Various platforms, including user-generated content sites like DeviantArt and Reddit, or publications like the *New York Times*, have opposed such practices. There is a shared sense of an exploitative dynamic at play: a commons of publicly available knowledge and culture being used as a raw material for commercial services that may capture its value without giving back. Without proper regulation protecting the digital commons, digital content will be exploited as AI systems grow, and as data are siphoned away into closed models by companies unwilling to support original content creation. The approach adopted by the EU balances the freedom to 'mine' content for the purpose of AI training with an opt-out mechanism available for those who want to reserve their rights.

Text- and data-mining rules should strike a balance between allowing content to be reused freely and protecting intellectual property. Measures enabling content owners to opt out of allowing their data to train AI systems are important for ensuring this balance.[86] But copyright rules themselves are not enough. A new social contract is needed to ensure that the profits generated by AI systems are recycled into funding production of the very content on which such systems rely; a financial levy might be one way of achieving this.[87]

Thirdly, **greater involvement of the public sector and increased public investment are needed to secure public interest in responsible AI**. Public research institutions and supercomputing centres are among the few actors that can compete with the commercial AI giants when it comes to research funding and computing infrastructure. In July 2023, the French government announced

86 Keller, P. (2023), 'Protecting creatives or impeding progress?', (2023), Open Future blog, 17 February 2023, https://openfuture.eu/blog/protecting-creatives-or-impeding-progress.
87 Keller, P. (2023), 'AI, the commons, and the limits of copyright', Open Future blog, 22 June 2023, https://openfuture.eu/blog/ai-the-commons-and-the-limits-of-copyright.

the outlines of a national AI initiative based on open-source principles, aimed at creating new models developed by national AI champions and complemented by publicly accessible training datasets.[88] In the UK, the Labour Party has proposed to increase tenfold the budget of the Foundation Models Taskforce in order to build BritGPT, to ensure that there is publicly owned capacity to develop and run foundation models.[89] In the US, leading AI researchers have argued for 'public option AI' and the need to provide substantial funding to the National Artificial Intelligence Research Resource – a pilot scheme launched by the U.S. National Science Foundation – so that it would offer public alternatives for computational power and data sources.[90] Finally, the EU announced in January 2024 the creation of ALT-EDIC, a consortium tasked with creating publicly available language resources for training language models.[91]

We are facing a challenging public debate, in which opinion leaders wield tremendous influence by virtue of also often being the owners of companies that are quickly concentrating power around the new AI technologies. These voices suggest that a safe AI future depends on societies trusting Big Tech to be gatekeepers of technologies that are complex, powerful, supposedly even sentient in a predictable future. And the debate itself may even be prone to centralization, as some parts of governments are willing to treat the new AI giants as the only voices they need to consult. This is a vision of technological development that is not in line with democratic values. And the fear of AI risks – most of these fears uncertain, extrapolated and exaggerated – are used to cement this concentration of power.

Instead, we need an approach that is democratic, with technologies serving citizens, and being available for use by citizens in ways that are affordable and just. The open-source approach, while not without its challenges, offers one of the clearest paths to attaining these goals, especially when coupled with a strong public commitment to developing AI as public infrastructure.

---

88 Chatterjee, M. and Volpicelli, G. (2023), 'France bets big on open-source AI', *Politico*, 4 August 2023, https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big.
89 Bayfield, H. (2023), 'Great British Cloud and BritGPT: the UK's AI Industrial Strategy Must Play to Our Strengths', Labour for the Long Term, 20 May 2023, https://www.labourlongterm.org/briefings/great-british-cloud-and-britgpt-the-uks-ai-industrial-strategy-must-play-to-our-strengths.
90 U.S. National Science Foundation (undated), 'National Artificial Intelligence Research Resource Pilot', https://new.nsf.gov/focus-areas/artificial-intelligence/nairr; Baksh, M. (2023), 'Leading Public-Interest Technologist Sees National Research Resource as a Potential Foundation for an "AI Public Option"', Schneier on Security, 1 December 2023, https://www.schneier.com/news/archives/2023/12/leading-public-interest-technologist-sees-national-research-resource-as-a-potential-foundation-for-an-ai-public-option.html.
91 European Commission (2024), 'Communication on boosting startups and innovation in trustworthy artificial intelligence', 24 January 2024, https://digital-strategy.ec.europa.eu/en/library/communication-boosting-startups-and-innovation-trustworthy-artificial-intelligence.

06
Resisting colonialism – why AI systems must embed the values of the historically oppressed

If dominated by major powers, AI development risks creating a new form of digital colonialism, particularly in Africa and other parts of the Global South. But a more optimistic future is imaginable, in which universal rules on AI are jointly shaped in a global public sphere drawing on many cultures and value systems.

Arthur Gwagwa

American and Chinese artificial intelligence (AI) systems, both their algorithms and data infrastructures, are in a contest for supremacy, which many in the AI and policy communities are following with interest. But for many countries around the world, the question of which model will prevail is secondary to the uncomfortable fact that both represent a similar force of foreign technology. AI imposed from outside, and shaped by the language and social systems of a few powerful countries, risks becoming a form of digital colonialism that ignores diversity of geography, language and culture.

How can today's post-colonial societies, such as many in Africa, avoid being recolonized, this time through foreign technology?[92] AI systems are not neutral intermediaries. They, like every technology, are tools of political power,[93] and attention should be paid not just to their technical implications but to their potential to disrupt and fragment societies in the same manner that colonialism

92 Gwagwa, A. and Townsend, B. (2023), 'Re-imagining Africa's sovereignty in a digitally interdependent world', *Global Policy*, 10 May 2023, https://www.globalpolicyjournal.com/blog/10/05/2023/re-imagining-africas-sovereignty-digitally-interdependent-world.
93 Helberger, N., Kleinen-Von Koenigsloew, K. and van der Noll, R. (2015), 'Regulating the New Information Intermediaries as Gatekeepers of Information Diversity', *Info*, Vol. 17 No. 6, 2015, pp. 50–71, https://ssrn.com/abstract=2728718.

did in analogue contexts in the past.[94] Colonial administrations in the 20th century sought to suppress the use of African native languages, and were especially anxious to promote the written use of European languages. 'People were taught to feel ashamed for their own language,' observes a researcher in natural language processing (NLP) – a branch of AI associated with enabling computers to understand text – quoted in one academic article.[95]

This risk of exclusion may even be accentuated by the inception of foundation and generative models in AI, for example if large language models (LLMs) such as ChatGPT rely on geographically, culturally or linguistically non-diverse sources: 'LLMs model their output on the texts they have been trained on, which is more or less the writing of the entire Internet, including all the biases – the prejudices, racisms, and sexisms – that constitute much of it … in the future, language models themselves may take on the status of a surrogate public sphere.'[96]

Diversity and inclusivity in AI need to be enshrined within, and supported by, an internationally developed human rights framework.

Viewing AI as a sociotechnical system[97] – not just as a tool – brings the values underlying AI to the surface. And it is essential that such values are diverse: representative not only of people on the West coast of the US, but also of other nationalities and cultures as well as the formerly oppressed, including women and thought leaders from the Global South. Diversity and inclusivity in AI need to be enshrined within, and supported by, an internationally developed human rights framework. This also means taking a critical eye to current imbalances in the global development of AI – such as the frequent marginalization of non-Western voices – and recognizing the problem's sources in institutional structures and historical inequalities.[98]

One way to think about addressing the issue is through the pursuit of what has been described as an 'overlapping consensus'[99] – one that would draw its values substantially from the Global South and Europe rather than just the US or China, and could thus inform a global AI-enabled ecosystem that is more equitable than one that excludes some parts of the world in its design.

---

94 Chan, A. et al. (2023), 'Harms from Increasingly Agentic Algorithmic Systems', FAccT '23, 12–15 June 2023, https://arxiv.org/abs/2302.10329.
95 Ravidran, S. (2023), 'AI often mangles African languages. Local scientists and volunteers are taking it back to school', *Science*, 20 July 2023, https://www.science.org/content/article/ai-often-mangles-african-languages-local-scientists-and-volunteers-are-taking-it-back.
96 Bajohr, H. (2023), 'Whoever Controls Language Models Controls Politics', Hannes Bajohr blog, 8 April 2023, https://hannesbajohr.de/en/2023/04/08/whoever-controls-language-models-controls-politics.
97 Van de Poel, I. and Kroes, P. (2014), 'Can technology embody values?', in Kroes, P. and Verbeek, P. (eds) (2014), *The Moral Status of Technical Artifacts*, pp. 103–24, Dordrecht: Springer, https://link.springer.com/chapter/10.1007/978-94-007-7914-3_7.
98 Michael Running Wolf, https://indigenousinai.org.
99 Hutson, M. (2023), 'Rules to keep AI in check: nations carve different paths for tech regulation', *Nature*, Vol. 620(7973), pp. 260–63, https://www.nature.com/articles/d41586-023-02491-y.

Values in technology and regulation

For the past two decades, values have been globally exported through digital technology. In the case of social media, for instance, norms around freedom of expression, newsworthiness and privacy have been renegotiated through the algorithms built in Silicon Valley. The new generation of AI tools are no different, with the risks associated with their influence increasing the more frequently we outsource our decisions to the human-like answers the tools might give.

In Africa, this means AI tools may often ignore African values and reflect those of the countries leading AI development, most notably the US and China. In very simple terms, US systems tend to emphasize the autonomy of the individual and commodify social relationships. Chinese systems, meanwhile, advance the value of social control.[100] To date, African countries have often had to choose between these two competing blueprints, even though neither necessarily benefits local cultures or provides a public good. For example, where governments such as Senegal have seemed to embrace the Chinese model of digital sovereignty (for example, by localizing government data on to domestic servers), such action has sometimes given the impression of performative policymaking for political ends. This may ultimately lead to increased, rather than reduced, hegemony of imported values along with a strengthening of foreign economic interests (Senegal's new national data centre, opened in 2021, was Chinese-built).[101]

Europe is a different case, in some ways less obviously invasive as an AI power, but also emblematic of the challenges and uncomfortable dilemmas African countries face as they seek to navigate the AI landscape and shape it to their advantage in the future. What Europe lacks in tech export capacity it makes up for in its world-leading regulation. In an attempt to boost member states' technological autonomy, and insulate European citizens from US and Chinese AI tools, the EU is developing an AI regulatory framework that will include protections for individuals, markets and digital products.[102] The most notable element of this initiative is the new EU AI Act, approved by the Council of the EU in May 2024.[103] (See also, in particular, Chapter 3, 'Regulating AI and digital technologies – what the new Council of Europe convention can contribute', and Chapter 5, 'Open source and the democratization of AI'.) The EU's global influence has given rise to anticipation, at least in Europe, that the world (including Africa) will embrace the standards enshrined in the AI Act, including those around values such as privacy and autonomy of the individual.

---

100 Townsend, B. and Gwagwa, A. (2023), 'Authoritarian Alliances and the Politicking of Data in Africa', *Journal of Online Trust and Safety*, Vol. 2, No. 1, 21 September 2023, https://www.tsjournal.org/index.php/jots/article/view/111.
101 Gwagwa and Townsend (2023), 'Re-imagining Africa's sovereignty in a digitally interdependent world'.
102 Christakis, T. (2020), "European Digital Sovereignty': Successfully Navigating Between the 'Brussels Effect' and Europe's Quest for Strategic Autonomy', Multidisciplinary Institute on Artificial Intelligence/Grenoble Alpes Data Institute, e-book, 18 December 2020, https://ssrn.com/abstract=3748098.
103 Council of the EU (2024), 'Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI', press release, 21 May 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai; Chee, F. Y. (2024), 'EU lawmakers ratify political deal on artificial intelligence rules', Reuters, 13 February 2024, https://www.reuters.com/technology/eu-lawmakers-back-political-deal-artificial-intelligence-rules-2024-02-13.

Yet in its own way, the EU's ostensibly progressive approach is also an unwelcome imposition of values on non-Western countries, and a form of domination based on paternalism. It means Africans, for example, may potentially be denied the right to choose to govern their societies based on their own values of community and the equitable distribution of social goods. Although it does not appear that African countries will be coerced into adopting EU regulations, in practice states may nonetheless choose to comply with the EU AI Act in order to access European markets – in much the same way as some African states have already adopted European cyber governance standards. In short, the regulatory power asymmetry between Europe and Africa that is partly a historical legacy may come into play again where AI regulation is concerned.

While not all European values are bad *per se*, the imposition of the values of individualism that accompany Western-developed AI and its regulations may not be suitable in communities that value communal approaches. Just as dual-use biometric technologies have the potential to create unintended consequences – for example, amplifying ethnic tensions[104] – the values that currently underpin AI deployment will likely lead to increased inequality alongside social, economic and political disruption, with technologically disadvantaged and under-represented populations in Africa faring the worst.[105]

The need for homegrown solutions

Given how AI systems may have disproportionately negative impacts on historically disadvantaged groups, more attention needs to be paid to how technology impacts the rights to self-determination in post-colonial societies. African societies have different approaches in this area from those of the countries and jurisdictions dominating the current discourse. For instance, on the question of whether humans owe ethical obligations to robots, African '*ubuntu*' values – which promote harmony, consensus, collective action and the common good – have thus far been excluded from the debate. Ethicists discussing the implications of robotics, in other words, have considered many variables but not how *ubuntu* fits into the picture.

Such dynamics confirm that we cannot expect solutions to come from the existing centres of power. The UK's proposed AI audits and the EU's comprehensive AI regulations are designed to protect European markets and preserve that continent's technological strategic autonomy and global dominance.[106] Paternalism can also be observed in how China collects African biometric data as a means to diversify AI training datasets that lack black faces; China needs datasets containing black faces to train the algorithms its AI labs produce. The promise of logistical efficiency that automation brings to the production and distribution of goods and services comes at the expense of African communal values. When Western companies harness

104 Townsend and Gwagwa (2023), 'Authoritarian Alliances and the Politicking of Data in Africa'.
105 Smith, M. L. and Neupane, S. (2018), *Artificial intelligence and human development: Toward a research agenda*, white paper, International Development Research Centre, https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949.
106 Christakis (2020), "European Digital Sovereignty': Successfully Navigating Between the 'Brussels Effect' and Europe's Quest for Strategic Autonomy'.

machine learning to improve the productive efficiency of industrial agriculture, they disrupt traditional societal structures that make African life meaningful. In addition, in globalized economic systems, major decisions on resource allocation are taken far from individual producers and consumers and have become opaque to them. These and other cases of value imposition by global AI superpowers show that Africa is 'a theatre of operations rather than the focus itself'.[107] When foreign values compete in this geopolitical theatre, they erode African collective values such as communalism. Despite having their own downsides, these values give meaning to Africans and support the political agency necessary to counteract external domination.

To secure such agency, multi-stakeholder approaches to AI governance are critical. Drawing from the readily available wealth of scholarship and expertise on resisting colonialism among the formerly oppressed,[108] such approaches will need to challenge fundamental assumptions about proprietary research; they will also need to address issues such as lack of representation, and the absence of mechanisms for shared ownership.[109] This is important given that AI governance discussions that only include regulators and tech companies miss critical voices: individuals and communities who are most affected by the vulnerabilities AI could create. The decision to listen, learn and invite new leaders to the table could shape an AI-driven future of equity, compassion, human creativity and opportunity, rather than one of exclusion and exploitation.[110]

The decision to listen, learn and invite new leaders to the table could shape an AI-driven future of equity, compassion, human creativity and opportunity, rather than one of exclusion and exploitation.

An inclusive AI partly informed by *ubuntu* values would work both ways: not only benefiting Africa but also providing normative standards for the rest of the world, to the benefit of all. In this more equitable digital world, European regulators would not be alone in pushing back against US and Chinese hegemony; they would have the support of the Global South. For this to occur, there needs to be a global public sphere in which universal rules on AI can be debated and forged. While such a sphere would certainly include and respect European voices, its heart might lie in the southern hemisphere, with the debate led by the perspectives of communities that have historically faced oppression and colonialism.[111]

---

107 De Carvalho, G. and Rubidge, L. (2022), 'Global geopolitical competition hits Africa: Can it maintain its voice?', ACCORD, 22 September 2022, https://www.accord.org.za/analysis/global-geopolitical-competition-hits-africa-can-it-maintain-its-voice.
108 Boscarino, N. (2023), 'AI Panic is Baby's First Colonialism', Nima Boscarino blog, 13 July 2023, https://acsweb.ucsd.edu/~nboscarino/blog/2023/ai-colonialism.
109 Van de Poel and Kroes (2014), 'Can technology embody values?'.
110 Bajohr (2023), 'Whoever Controls Language Models Controls Politics'.
111 Barbrook, R. and Cameron, A. (1995), 'The Californian ideology', *Mute*, Vol. 1, No. 3, 1 September 1995, http://www.metamute.org/editorial/articles/californian-ideology.

In this way, the above-mentioned 'overlapping consensus'[112] could bring together the best thinking from the Global South and Europe to create a safer, more sustainable and more equitable vision for the future of AI. Such a consensus, based on an intercultural discourse, can ultimately address the unfair distribution of benefits and harms of AI by evaluating the systemic colonial social power arrangements behind such a distribution.

In taking this approach, we will build better AI, too: systems that spotlight historical inequalities and locate problems not just within technical systems, but within the social structures and institutions they originate from.[113]

112 Hutson (2023), 'Rules to keep AI in check: nations carve different paths for tech regulation'.
113 Michael Running Wolf, https://indigenousinai.org.

07
The UK needs a 'British AI Corporation', modelled on the BBC

UK policy responses to AI have focused on promoting private sector innovation. But widespread growth from AI is unlikely until it has earned the public's trust. To build AI systems that strengthen fairness, honesty and creativity across the UK, a new public-service AI institution is needed – in short, a kind of BBC for AI.

Brandon Jackson

We are told the fourth industrial revolution is here, and that the UK is on the front foot. Since the launch of the UK's National AI Strategy in 2021, the promise of artificial intelligence (AI) to unleash 'productivity, growth and innovation across the private and public sectors' has been a common political refrain.[114] AI seems to be everywhere – except, for now, in the productivity statistics.[115] Techno-optimistic soundbites from government and industry have done little to improve the UK's public mood around technology – a mood subdued by warnings of job losses, by injustices like the Horizon IT scandal, and even by fears about the possibility of AI-driven extinction. One recent poll showed that only 18 per cent of British people are optimistic about AI.[116]

AI's trust problem must be addressed. UK policymakers must recognize that the link between AI inventions and productivity growth is not automatic. Instead, history shows that such growth will only occur when the public trusts technologies enough to adopt them deeply into daily economic life.[117] That's why the UK needs

114 UK Department for Science, Innovation and Technology (2021), *National AI Strategy*, 22 September 2021, https://www.gov.uk/government/publications/national-ai-strategy.
115 Office for National Statistics (2024), 'Productivity flash estimate and overview, UK: October to December 2023 and July to September', 15 February 2024, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/ukproductivityintroduction/octobertodecember2023andjulytoseptember2023.
116 Smith, M. (2023), 'Britons lack confidence that AI can be developed and regulated responsibly', YouGov, 1 November 2023, https://yougov.co.uk/technology/articles/47744-britons-lack-confidence-that-ai-can-be-developed-and-regulated-responsibly.
117 Gordon, R. J. (2016), *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*, Princeton Economic History of the Western World series, Princeton: Princeton University Press, https://press.princeton.edu/books/hardcover/9780691147727/the-rise-and-fall-of-american-growth.

a new public-service AI institution to help society navigate the technological changes ahead: a 'British AI Corporation' or BAIC, modelled roughly on the BBC. Such an institution could earn the public trust by building accountable AI systems that help solve important problems, powered by a self-funding financial model that can sustain this essential 21st-century infrastructure indefinitely.

In search of a trustworthy partner

Early UK efforts to strengthen public trust in AI have been insufficient. This partly reflects the fact that the public sector has largely forfeited its role as a builder of trustworthy technologies in general, most often by outsourcing technical capabilities to the private sector. Despite the extraordinary public outcry after Fujitsu's failed Post Office computer systems, justice has been slow and the company's technology remains embedded in UK public infrastructure.[118] Meanwhile, a contract worth up to £330 million to build a data platform for the National Health Service (NHS) was recently awarded to a consortium of private companies led by US-based Palantir, in spite of opposition from the British Medical Association and the Doctors' Association UK, the latter of which called for work to be paused to 'ensure public trust, value for money, a trustworthy partner and patient consent'.[119]

The next wave of technologies is on track to be dominated by the same handful of international firms responsible for the last 20 years of consumer-facing digital technology.

At the same time, there is a growing sense that the private sector also cannot be relied on as the only route to trustworthy AI systems. The next wave of technologies is on track to be dominated by the same handful of international firms responsible for the last 20 years of consumer-facing digital technology. This concentration of market power is not just bad for competition. It also promises to create a lasting source of mistrust, as the gulf widens between the goals of the public and those of the private AI labs already locked into a race to secure market dominance by being the first to build human-level intelligences.[120]

---

118 Ungoed-Thomas, J. (2024), 'Fujitsu won £1.4bn in new government contracts after court ruling on Post Office software bugs', *Guardian*, 10 February 2024, https://www.theguardian.com/business/2024/feb/10/fujitsu-won-14bn-government-contracts-court-ruling-post-office-horizon-software-bugs.
119 Mann, A. (2023), 'Doctors call for pause in NHS Federated Data Platform contract', Doctors' Association UK, 11 November 2023, https://www.dauk.org/news/2023/11/11/doctors-call-for-pause-in-nhs-federated-date-platform-contract; and NHS England (2023), 'New NHS software to improve care for millions of patients', news release, 21 November 2023, https://www.england.nhs.uk/2023/11/new-nhs-software-to-improve-care-for-millions-of-patients.
120 Narechania, T. and Sitaraman, G. (2023), *An Antimonopoly Approach to Governing Artificial Intelligence*, Vanderbilt Policy Accelerator for Political Economy & Regulation, https://cdn.vanderbilt.edu/vu-URL/wp-content/uploads/sites/412/2023/10/06212048/Narechania-Sitaraman-Antimonopoly-AI-2023.10.6.pdf.pdf.

Thus far, the UK government's strategy has been to navigate this tension by creating innovative regulations that will make these 'frontier AI systems' safer.[121] But policymakers have not engaged at all with the primary concerns of the public – namely, the degree to which AI will affect jobs and society. As a result, polling shows that only 18 per cent of the British public trust tech companies to build AI responsibly, and that only 14 per cent trust the government to regulate it responsibly.[122] There is little reason to believe the situation will improve. That is why a new approach is needed.

One chapter in British history illustrates how transformative technologies can both drive growth and increase trust if a public option is empowered to lead the way. After the First World War, Britain looked nervously across the Atlantic as the new technology of the day – radio – became an overnight commercial success in the US. The UK government was faced with a dilemma. Radio offered the prospect of supercharging growth in the nascent domestic electronics manufacturing sector. Yet early radio culture was seen as a dangerous outgrowth of American capitalism, powered by machines resembling scientific apparatus that no one wanted to bring into their homes.[123]

It was at this moment that the BBC was founded in 1922 to find a way to drive technological adoption by balancing the need for growth with the need to protect British values. These goals were built into the BBC's institutional design. Its original funding was directly tied to growth: it took a cut of the income from every radio set sold. This meant the BBC had to work hard to invent usages of the new technology that the British people would actually want. In the early days, as the BBC's first director-general, John Reith, put it, 'Few knew what they wanted, fewer what they needed.' That's why the BBC decided to go further, aiming 'to carry into the greatest number of homes everything that was best in every department of human knowledge, endeavour and achievement; and to avoid whatever was or might be hurtful'.[124] This ambition to meet public needs was soon enshrined in a mission to 'inform, educate and entertain'.[125]

The bet to build a public broadcaster paid off. Technological adoption and manufacturing growth were swift. Yet only in the long run has the true impact become clearer. Just as railways connected the regions of the UK in the 19th century, the BBC became a key part of the infrastructure of 20th-century British life, dependably connecting citizens with the arts, the state, and the truth.

These outcomes stand in sharp contrast to the bitter experiences of recent infrastructural history. With examples ranging from misinformation on social media to sewage in our rivers, we have seen the dangers that arise when private interests control the networks that connect us. That is why the time is right to invest in a trustworthy new AI partner that operates in the public interest.

---

121 UK Department for Science, Innovation and Technology (2021), *National AI Strategy*.
122 Smith (2023), 'Britons lack confidence that AI can be developed and regulated responsibly'.
123 Briggs, A. (1985), *The BBC: The First Fifty Years*, Oxford: Oxford University Press.
124 Reith, J. (1949), *Into the Wind*, London: Hodder & Stoughton.
125 BBC (undated), 'Mission, values and public purposes', https://www.bbc.com/aboutthebbc/governance/mission.

A British AI Corporation

Drawing upon the UK's rich history of innovative public infrastructure, the government should establish a public option for AI by creating a new 'British AI Corporation' – a BAIC rather than a BBC, as it were.[126] This new institution would ensure that everyone has access to powerful, responsibly built AI capabilities. Yet the BAIC should be more than just a head-to-head competitor with the private AI companies.[127] It should be set up with an institutional design that empowers it to chart an independent path, building innovative digital infrastructure in the public interest.

The BAIC should be founded with a clear charter to which it can be held accountable. At the heart of this charter must be a mission with the clarity and timeliness of the BBC's: to build AI systems in the public interest that strengthen fairness, honesty and creativity throughout the UK. This mission would ensure that rather than focusing on the most profitable or amusing use cases, the BAIC would be compelled to address problems that matter most to the British public: shaping a fairer society rather than increasing inequalities, amplifying the truth and not misinformation (or disinformation), and ensuring that AI empowers artists rather than automating away creativity.

Achieving these goals will require creativity in turn. That's why a charter must be complemented with seed funding to ensure the institution's independence. Such funding would grant the BAIC time and space to experiment, fail and learn, as start-ups in the private sector are often able to do. To reduce costs, the BAIC should be given preferential access to the new public computing infrastructure being built across the UK, conditional on the new institution complying with its mission.[128] With that support, a modest initial investment in the order of £250 million would immediately make the BAIC one of the largest players in the London tech scene, allowing it to hire hundreds of experts at market rates for several years to bring world-class AI systems to market.[129]

Lastly, the new institution must be fully incentivized to build AI systems that the public wants to use. This can be achieved via a financial model predominantly underpinned by AI product usage rather than by state funding. This growth-based funding was one of the secrets of the BBC's early success. Depending for revenue on the public uptake of BAIC-developed AI systems would focus minds on removing barriers to adoption. Such a model would generate revenues that both finance the institution's operations and allow it to invest in future innovation.

---

126 One of the first orders of business must be to devise a better name than BAIC.
127 For the economic benefits of a public option, see Coyle, D. (2022), 'The Public Option', *Royal Institute of Philosophy Supplements*, 91 (May 2022): 39–52, https://doi.org/10.1017/S1358246121000394. For the national security benefits, see Belfield, H. (2023), 'Great British Cloud and BritGPT', Labour for the Long Term, 20 May 2023, https://www.labourlongterm.org/briefings/great-british-cloud-and-britgpt-the-uks-ai-industrial-strategy-must-play-to-our-strengths.
128 University of Bristol (2023), 'Unprecedented £225m investment to create UK's most powerful supercomputer in Bristol', press release, 1 November 2023, https://www.bristol.ac.uk/news/2023/november/supercomputer-announcement.html.
129 For example, £250 million would fund paying a team of 500 people an average annual salary of £100,000 for five years.

Building trustworthy infrastructure

With these institutional design features setting it up for success, the BAIC could start solving problems that matter most by building trustworthy AI infrastructure that includes all of society. First, a new BAIC should tackle AI's trust problem by becoming a model for a publicly accountable tech company. Many currently opaque decisions throughout the AI development process could and should be subjected to public input and scrutiny: ranging from deciding which problems to solve at the start of the planning process all the way 'downstream' to ensuring that outputs generated by AI systems align with shared values.[130] The BAIC will inevitably get some things wrong. When this happens, it must admit its mistakes, change direction and embrace its accountability to the public.

Trust could also be earned by building AI systems in a fair way that strengthens the commons rather than enclosing it. While many profit-driven tech companies say they want to put the best in every department of human knowledge into users' hands, few are willing to pay for this. Instead, they are often incentivized to 'free-ride' by scraping the public domain for data without permission. The BAIC could and should play fairly. It could partner with cultural institutions that have the vital role of maintaining and expanding our collective knowledge, and voluntarily pay for access to their 'data troves'.[131] This public-interest data infrastructure would ensure that, as AI grows, so too would the funding available to the institutions working hard to steward the UK public's inherited commons responsibly.

Public-interest data infrastructure would ensure that, as AI grows, so too would the funding available to the institutions working hard to steward the UK public's inherited commons responsibly.

Second, the BAIC should tackle the UK's productivity growth problem by doing more to address the concerns of businesses and employees. At present, the AI sector is extremely difficult for most companies and organizations to influence, much less compete in, due to the presence of a few well-funded incumbents focusing primarily on their own market share. This makes it harder to achieve bottom-up growth that could drive productivity. A new BAIC could build infrastructure that the Competition and Markets Authority has identified as having the potential to lower barriers to entry, such as systems to make it easier to switch models and datasets across platforms.[132] This work might do for AI what Channel 4 did for broadcasting, by helping to create a new ecosystem in which small firms can build world-class services and share them with the world.

---

130 The Collective Intelligence Project (2024), *A Roadmap to Democratic AI*, March 2024, https://cip.org/research/ai-roadmap.
131 Serpentine Arts Technologies (2024), 'Future Art Ecosystems 4: Art x Public AI', March 2024, https://reader.futureartecosystems.org/briefing/fae4/introduction.
132 Competition and Markets Authority (2023), 'AI Foundation Models: Initial report', 18 September 2023, https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

The BAIC could also stimulate growth by building AI that makes the workplace fairer, addressing employee concerns that AI will spark a 'job apocalypse'.[133] The public are understandably concerned that disruption will occur as new technologies develop faster than policymakers can regulate them effectively. As an AI developer accountable to the public for making life in the UK fairer, the BAIC would offer an extra layer of protection. The BAIC could set out a principled vision for ethical automation, and then lead the way by translating this vision into useful products. The BAIC could innovate by prioritizing capabilities that complement rather than replace labour.[134] And if certain functionality is found to decrease fairness, an independent BAIC could price it differently to reduce its impact or could simply switch it off. These safeguards could reduce resistance to automation and spark responsible productivity growth.

Finally, a BAIC could help ease the public sector's dependence on private contractors. Although there have been promising early signs of the British civil service investing in the recruitment of in-house AI talent, the direction of travel is still very much towards contracting with private AI platforms to design and supply government systems. This not only creates privacy risks and threatens to be poor value for money – it also represents a missed opportunity to inspire public sector innovation. Instead, a BAIC could identify the core feature sets required by both civil servants and the public for a given AI application, and start competing for the relevant contracts. The resulting product would become shared infrastructure owned by the public. This would transform procurement from a process in which millions (or billions) of pounds disappear into private sector contracts into a transparent process of public investment – giving taxpayers a better deal and establishing a pathway towards long-term financial sustainability for the new BAIC.

With a roadmap deeply aligned with their own values and priorities, members of the British public could stop worrying about AI, and simply get on with finding clever ways to integrate it into their lives.

---

133 Jung, C. and Desikan, B. (2024), *Transformed by AI: How generative artificial intelligence could affect work in the UK – and how to manage it*, Institute for Public Policy Research, March 2024, https://www.ippr.org/articles/transformed-by-ai.
134 Acemoglu, D. and Restrepo, P. (2018), 'The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment', *American Economic Review*, Vol. 108, No. 6 (June 2018): pp. 1488–1542, https://doi.org/10.1257/aer.20160696.

08
An ethics framework for the AI-generated future

The autonomous capabilities of emerging AI systems pose societal risks if consequential decisions are based on flawed information and are unguided by appropriate ethical parameters. This essay proposes a process for determining the level of oversight – informed by ethical considerations – needed for safer AI.

Micaela Mantegna

This essay was 100 per cent written by a human being.

Not long ago, starting an essay with this proposition might have been seen as very odd. Today, the fact that this seems an increasingly reasonable warranty is testament to the meteoric rise of generative artificial intelligence (GAI). In this new paradigm, AI models are able to create information by learning how to mimic the data in their training datasets, generating convincing examples of photos, computer code, music – or essays like this one.

As humans, our learning processes and definitions of truth and reality are connected to empirical observation. For decades, computational models have manipulated who sees what. With its newfound generative capabilities, GAI is taking this manipulation to the next level. When creating synthetic data, GAI is not just redefining creativity but also convincingly challenging our natural trust in the information we perceive through our senses. Is the politician's speech that is going viral on social media real or fake? Am I talking to a human customer representative, or is a polite bot replying to my chats? How can we be sure what is real, ever again?

At present, the implications of GAI for the future of democracy, the economy, labour and education are unfathomable, in terms of both the thrilling possibilities and chilling risks. This essay explores some of the new challenges that this generative future presents to the already-complex landscape of AI ethics, underscoring how critical it is for society to reach timely agreement on policy frameworks for ethical innovation.

AI garbage in, AI garbage out

Despite the attention it is now attracting, the field of AI is not new. It has been around for decades, and has seen an evolution in approaches and techniques. The latest advancements arise in part from developments in machine learning models, enabled by the conjunction of data availability (thanks to the ubiquity of the internet and smart devices), mathematical research, increasing computing power, and lower costs for data storage and retrieval.

While previous approaches such as expert systems were about codifying logic and rules of knowledge, machine learning can be thought of as 'learning by example'. When you have seen enough pictures of cats, you can recognize one in new information, even if you cannot explain the rationale behind how you know it to be a cat.

At its core, an AI model is a stack of algorithms. An algorithm is a sequence of instructions to transform an input into an output. In a way, we can think of it as a recipe: ingredients (input) are processed according to prescribed steps to achieve a result (output). Crucially, the final product depends both on the quality of the ingredients and on using the right recipe. It is impossible to bake an apple pie if bananas are the input, or if the recipe was created by a system trained with oranges. The same goes for AI models: biased data lead to biased results. If garbage goes in, garbage comes out.

Machine learning makes AI models extremely dependent on data collection, which in the context of our current internet landscape has contributed to the rise of 'surveillance capitalism'.

Machine learning makes AI models extremely dependent on data collection, which in the context of our current internet landscape has contributed to the rise of what Shoshana Zuboff has called 'surveillance capitalism'.[135] Targeted advertising has become the lifeblood of digital economies, shaping both how digital platforms and its products are designed and how we interact with them.

Before the current wave of GAI, AI models were focused on prediction and classification tasks, like recommendation engines suggesting the next video to watch, or which products might be appealing for specific consumers based on their purchase history. More worryingly, such systems started to be deployed to predict future behaviours in delicate and highly contextual matters, such as the likelihood of someone defaulting on a mortgage payment or committing a crime.

These types of statistical systems look into the past to try to predict the future. They extract features from data to build a model that interprets the world and aims to predict future outcomes to similar problems. While mathematically sound,

135 Zuboff, S. (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile.

this proposition has a fundamental problem: as they aim to capture and describe a situation in the real world, AI models replicate the biases and inequalities in the reality that they purport to observe. Consequently, those biases and inequalities are perpetuated in the outputs and projected into the very same future AI tries to predict. This creates a vicious cycle and makes AI systems conservative and risk-averse, biased towards the status quo. It also risks history repeating itself: if data show that some jobs, roles and industries have been male-dominated, a model developed to extract the best candidates in the same fields will tend to pick people with similar profiles. Early recruitment models for traditionally male-dominated roles, for instance, discriminated against female applicants.

Furthermore, even if AI models can simulate intelligent results, they lack contextual awareness and common sense, which makes them unsuitable for dealing with nuanced linguistic tasks, pondering values or moderating content.

Foreseeable challenges posed by GAI: beauty, truth, hallucinations and anthropomorphization

Understanding AI blind spots and ethical problems is critical because, despite mitigation strategies and technical safeguards,[136] those flaws are being carried on into GAI and will be true for whatever AI breakthroughs come next.

Historically, dealing with bias has been challenging for AI. This is especially the case for generative models because the biases embedded in the content they produce create a new layer of digital reality in terms of visual languages or factoids. By reflecting reality through distorted lenses, and releasing back into the world content according to that point of view, GAI is creating ontological aesthetics and semantics, producing new cultural signifiers. As the internet becomes flooded with synthetic content, the stereotypes, misconceptions and falsehoods produced by AI systems spill over into people's actual beliefs and perceptions, effectively becoming 'real' as they are assimilated by society. This creates another vicious cycle, as today's synthetic content becomes tomorrow's training data, perpetuating bias into the future. For example, GAI image generators prompted to create an image of a doctor are likely to produce an image of a male doctor. AI-generated images are also redefining ideals of beauty by defaulting to hegemonic, unattainable and synthetic standards of physical perfection[137] that risk exacerbating body dysmorphia rampant among vulnerable social media users.

AIs' blurring of fact with fiction is creating novel problems for users, including legal risks associated with reliance on AI-generated text and 'analysis' in situations where real-life accuracy is demanded. This is not necessarily the result of bad actors. Large language models (LLMs) are complex pieces of intellectual machinery that can fail, providing confident-sounding but spectacularly wrong answers.

---

136 Bianchi, F. et al. (2023), 'Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale', *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, June 2023*, pp. 1493–1504, https://doi.org/10.1145/3593013.3594095.
137 Llach, L. (2024), 'Meet Aitana, Spain's first AI model, who is earning up to €10,000 a month', euronews.next, 22 March 2024, https://www.euronews.com/next/2024/03/22/meet-the-first-spanish-ai-model-earning-up-to-10000-per-month.

In computer science, these faux pas have been christened AI 'hallucinations' or 'delusions' – instances in which AI models fabricate information entirely, while confidently behaving as if they are stating facts.[138]

One of the education challenges ahead is to strengthen humanity's critical thinking skills in relation to confidently presented errors or misrepresentations of fact.

This connects with another rising trend, unwarranted human reliance on AI systems as oracles and authoritative sources.[139] Just as people tend to trust confident-sounding speakers,[140] we should consider carefully the bond of 'epistemic trust' that is developing between humans and LLMs.[141] When one considers also how the replies provided by LLMs are generally detached from sources, one of the education challenges ahead is to strengthen humanity's critical thinking skills in relation to confidently presented errors or misrepresentations of fact.

GAI models are also susceptible to specific security risks. In a sort of AI hypnotic suggestion, attackers can manipulate the prompts given to AI models to induce forced answers. These 'prompt injections' can hijack and override safeguards,[142] poisoning the resultant output, which will be provided according to the new instructions, inadvertently to the user. Other vectors attack GAI models in a way similar to social engineering, by persuading or confusing the model, tricking it into providing answers or overriding safety guardrails.[143]

Considering how internet algorithms rank and position information according to relevance, if false information is disseminated and repeated enough by the right sources, it will find its way to the front page of search results. And from then,

---

138 In this, AI is similar to 'Cantinflas', the iconic character created by Mexican comedian Mario Moreno famous for monologues full of incoherent and incorrect assertions. While boastful and shallow, they were so convincing that they conned the listener into thinking he was an expert. According to the Royal Academy of Spanish Language (RAE), *cantinflear* became a term to refer to someone speaking or acting in a nonsensical and incongruous manner and saying nothing of substance. As nobody would use a doubtful AI system, there is a design incentive to provide answers, which in turn conflated with hallucinations leads to generative language models often suffering from acute cases of Cantinflas syndrome. See also https://www.youtube.com/watch?v=Y7quI7z63dE.

139 Just as in past decades the term 'Google it' became synonymous with searching, now we increasingly hear people nonchalantly saying, 'I asked ChatGPT.'

140 Pulford, B. D., Colman, A. M., Baubang, E. K. and Krockow, E. M. (2018), 'The Persuasive Power of Knowledge: Testing the Confidence Heuristic', *Journal of Experimental Psychology: General*, 147(10), pp. 1431–44, https://doi.org/10.1037/xge0000471.

141 Kushnir, T., Sobel, D. and Sabbagh, M. (2022), 'Trust comes when you admit what you don't know – lessons from child development research', The Conversation, 15 February 2022, https://theconversation.com/trust-comes-when-you-admit-what-you-dont-know-lessons-from-child-development-research-175596.

142 Paradoxically, while there is so much praise of, and debate over, the 'intelligence' of these conversational AI systems, one blind spot in terms of cybersecurity is their candid display of naivety. See Binder, M. (2023), 'ChatGPT plugins face 'prompt injection' risk from third-parties', Mashable, 27 May 2023, https://mashable.com/article/beware-chatgpt-ai-prompt-injections; and Burgess, M. (2023), 'The Security Hole at the Heart of ChatGPT and Bing', *Wired*, 25 May 2023, https://www.wired.com/story/chatgpt-prompt-injection-attack-security.
On a lighter note, AIs perform better in maths when asked to reply as a Star Trek character: Guenot, M. (2024), 'AIs are more accurate at math if you ask them to respond as if they are a Star Trek character – and we're not sure why', *Business Insider*, 29 February 2024, https://www.businessinsider.com/using-star-trek-prompts-boost-ai-chatbot-basic-math-performance-2024-2.

143 Lakera, an AI security company, has developed a security challenge in which the user needs to convince or trick an AI guardian into divulging a password. See https://gandalf.lakera.ai.

it is a short data-mining step away from that information ending up in training datasets. As with propaganda, synthetic truths can crystallize into ideas systemically presented as real in a sort of self-fulfilling prophecy.

Another looming problem is emotional manipulation and attachment. As GAI creates agents able to interact in real time in a way that can be tailored to specific users, there are already reported cases of emotional bonding of humans with AI bots.[144] Conversely, a journalist's conversation with Microsoft Bing's chatbot took a bizarre turn when the system declared its love for him and suggested that he break up with his wife.[145]

A related problem is anthropomorphization, in which human qualities, emotions or intentions are attributed to AI systems.[146] The conversational nature of people's interaction with LLMs, as well as the fact that AIs are sometimes embodied in human-looking robots or avatars, contributes to the confusion. Sometimes there are legitimate debates about the consciousness and personhood of AI systems, but more frequently than not the question is raised for shock value or the sake of marketing. What is certain, however, is that anthropomorphization can be a distraction from pressing and practical ethical concerns about the use of AI, and that this can contribute to public misrepresentation of AI's capabilities and limits.[147]

Understanding the different dimensions of AI ethics

Despite the challenges outlined above, the temptation to implement AI for the sake of automation is high. Public institutions are particularly suggestible to the promises of modernization, security and efficiency. AI systems are already being deployed in areas like justice and surveillance without proper safeguards, assessments, or possibilities for recourse in the event of error. Ethically,

---

144 This trend of 'attachment as a service' involves business models taking advantage of emotional bonds. See also Cole, S. (2023), "It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection', VICE, 15 February 2023, https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates; and The Project (2023), 'Replika ChatBot Users Devastated After AI Update Destroyed Their Relationship', 5 March 2023, https://www.youtube.com/watch?v=TmoWGV9IWsU.

145 Pringle, E. (2023), 'Microsoft's ChatGPT-powered Bing is becoming a pushy pick-up artist that wants you to leave your partner: 'You're married, but you're not happy'', *Fortune*, 17 February 2023, https://fortune.com/2023/02/17/microsoft-chatgpt-bing-romantic-love.

146 Dubois-Sage, M., Jacquet, B., Jamet, F. and Baratgin, J. (2023), 'We Do Not Anthropomorphize a Robot Based Only on Its Cover: Context Matters too!', *Applied Sciences* 13(15), 8743, 28 July 2023, https://doi.org/10.3390/app13158743.

147 Not long ago, the United Nations held a press conference with humanoid robots, including Sophia, the robot that in 2017 was named as the UN Development Programme's first Innovation Champion and also granted Saudi Arabian 'citizenship'. See AP News (2023), 'UN tech agency rolls out human-looking robots for questions at a Geneva news conference', 7 July 2023, https://apnews.com/article/humanoid-robots-better-leaders-ai-geneva-486bb2bad260454a28aaa51ea31580a6; and Center for International Communication (2017), 'Saudi Arabia Is First Country In The World To Grant A Robot Citizenship', press release, 25 October 2017, https://web.archive.org/web/20171110114747/https:/cic.org.sa/2017/10/saudi-arabia-is-first-country-in-the-world-to-grant-a-robot-citizenship. In the July 2023 press conference, journalists asked questions and the robots expressed their 'opinions', nonchalantly stating that they could eventually run the world more efficiently than humans, because they are unburdened by emotions and unbiased. AFP News Agency (2023), 'AI robots tell UN conference they could run the world | AFP', 7 July 2023, https://www.youtube.com/watch?v=cItRwEGThvo. Considering how these robots were powered by AI software and therefore replicating the very problems discussed in this essay, those PR stunts had very damaging effects. See also Hundt, A. et al. (2023), 'Robots Enact Malignant Stereotypes', *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea*, https://doi.org/10.1145/3531146.3533138.

determining when and where to implement automated decision-making systems and how they affect society is complex. Not every problem can or should be the subject of an automated solution, certainly not now and perhaps not ever.

Ethically, determining when and where to implement automated decision-making systems and how they affect society is complex. Not every problem can or should be the subject of an automated solution, certainly not now and perhaps not ever.

These are not decisions reserved solely for big corporations or governments. Today, everyone interacts with automated systems in different capacities, in the workplace, in public spaces or in private life. Some people may have room to choose to avoid AI tools, but others have no such choice.[148] There are, however, some principles that might inform a more responsible approach to the use of these models, and some considerations that should be taken into account. To help academics, policymakers or concerned citizens navigate this issue in an informed manner, this essay presents a framework that breaks these complex problems into a sequence of analytical steps that can be adapted as needed to different situations.

The first step is to create an ethical AI matrix[149] weighing the context, potential harms and potential gains so that a user can determine whether implementing an AI system instead of other solutions is justified, and the extent of automation that might be appropriate in a given case (this would cover a spectrum – from fully automated decision to requiring human oversight to unsuited for automation). To assess this, the matrix considers three vectors, outlined in Table 1.

148 For example, schoolteachers have found themselves on the front line here: dealing with essays written by LLMs, and having to decide whether such texts are human- or AI-made, while also starting to deploy tools for automatically grading them as well. Other workplaces have seen the use of AI tools either mandated by management, or banned altogether. See Mok, A. (2023), *Business Insider*, 'Amazon, Apple, and 12 other major companies that have restricted employees from using ChatGPT', 11 July 2023, https://www.businessinsider.com/chatgpt-companies-issued-bans-restrictions-openai-ai-amazon-apple-2023-7.
149 The author defines this as embodying the 'proportionality in context' principle. This means assessing how justified and proportional the rationale for automating a decision is, considering other options, gains and harms.

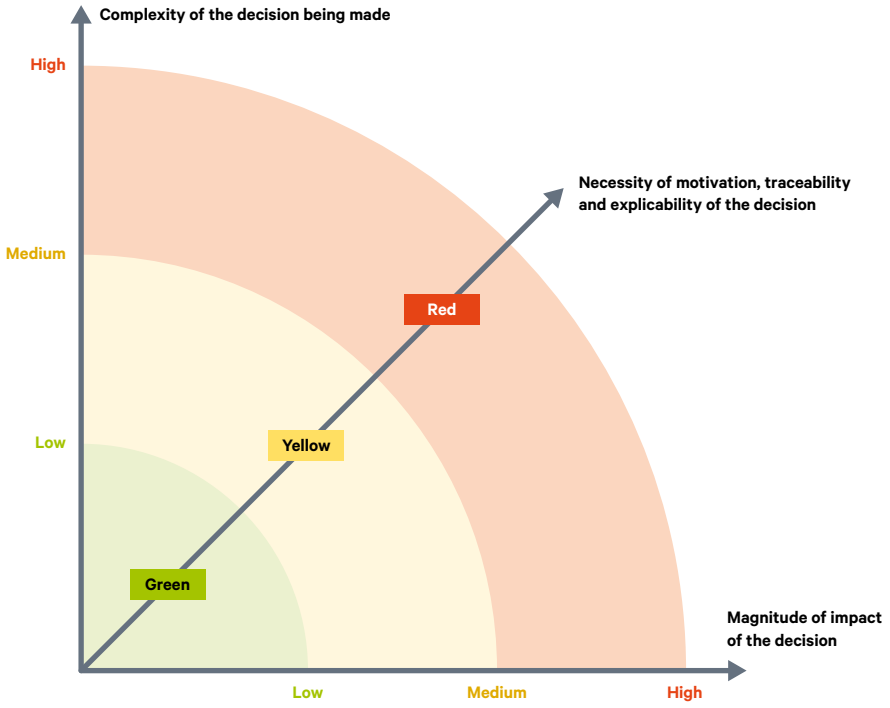
Table 1. Ethical AI matrix

Factor	Principles guiding assessment
a) Complexity of the decision being made	As examples from the previous sections show, AI is not good at understanding values and contextual references.[150] When evaluating whether or not to implement AI, it is worth recalling Andrew Ng's old but still applicable 'one second rule', meaning that a task is best suited for automation if a normal person could do it with less than one second of thought.[151] Is it a straightforward linear decision? Go ahead! Does the decision require considering values or understanding context? If so, while the speed and scale of AI present challenges, it may be wise to involve humans in some capacity: as decision-makers, evaluators or assessors.
b) Magnitude of impact of the decision	Second, we might consider how greatly a decision affects an individual or group of people. Is AI set to cause harm? Is the potential harm trivial or meaningful? Can that harm be undone or compensated for? Is there a reason why a person or group has to suffer that harm while others do not? Let's consider, for example, two recommendation engines: one for movies and the other for parole decisions. Even if conceptually both are AI-assisted decision-making systems, one that suggests a movie has a very different potential impact from one that recommends a parole decision. Decisions of no consequence should carry a different weight from those that could have meaningful, hard-to-reverse and expansive consequences.
c) Necessity of motivation, traceability and explicability of the decision	Lastly, we should ask how far we can understand and explain the decision made by an AI model. As AI tools become more sophisticated, they become less transparent. A system might be able to recognize a picture as a cat, but it probably cannot explain why it arrived at that conclusion. Considering how AI image recognition systems work in terms of statistical probability, an image with an 80 per cent probability of being a cat will likely be classified as such. For certain decisions, that level of reliability is not good enough, such as with facial recognition systems deployed by law enforcement agencies.[152] Measuring accuracies, testing edge cases, and understanding as far as possible how automated systems make decisions are critical tests for responsible AI deployment.
Determine course of action	
GREEN Implement automated decision	To determine if action in a given area could ethically be automated, the three factors above should be combined in a vector of magnitudes (low, medium, high) to establish green, yellow and red flags. For example: • A court verdict ranks highly in all three categories (i.e. complexity, magnitude, necessity/explicability): it is a very nuanced decision that requires critical assessment of facts, context, values, legal frameworks, etc.; it will have a significant impact on a person's life; and it needs a clear explanation from the court of the reasonings for the decision. • A film recommendation on a video-streaming system ranks green in all three categories. Depending on the viewer's past viewing history, the decision could be very straightforward. And the worst that could happen is that the viewer loses a couple of minutes watching a film that doesn't interest them. A detailed explanation for why the system arrived at the movie recommendation is not needed.
YELLOW Provide human oversight or recourse	
RED Never implement AI	

150 Shane, J. (2018), 'Do neural nets dream of electric sheep?', AI Weirdness blog, 2 March 2018, http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep.
151 Ng, A. (2016), 'What Artificial Intelligence Can and Can't Do Right Now', *Harvard Business Review*, 9 November 2016, https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now.
152 Democracy Now! (2023), 'False Arrest of Pregnant Woman in Detroit Highlights Racial Bias in Facial Recognition Technology', 9 August 2023, https://www.democracynow.org/2023/8/9/false_arrest_of_pregnant_woman_in; and Levin, S. (2018), 'Amazon face recognition falsely matches 28 lawmakers with mugshots, ACLU says', *Guardian*, 26 July 2018, https://www.theguardian.com/technology/2018/jul/26/amazon-facial-rekognition-congress-mugshots-aclu.

Visually, the ethical AI matrix can be constructed as in Figure 1:

Figure 1. Visual representation of proposed ethical AI matrix



Source: Author's illustration.

The second step involves understanding the different aspects of how AI impacts society, as outlined in Table 2.

Table 2. Assessing the collective impacts of AI on society according to eight factors

Awareness	Are users informed whether they are interacting with or subjected to an AI system? What solutions (such as adding warnings of interaction) are available to ensure awareness? Example: Do users know if they are speaking with a human or a chatbot?
Pervasiveness	Can we opt out from interacting with, or being subjected to, an AI system? Are there alternatives? Is there a real choice in terms of avoiding its use? Do the consequences of opting out make it too onerous or impossible to escape interactions with AI? Example: Facial recognition is being deployed in different airports in the US before passengers board their flights. Passengers are allowed to ask for an alternative identification method, but are unsure if that creates a problem with airport security; this results in a perverse incentive for passengers to accept such surveillance passively.
Scalability	Due to the scalable nature of software systems and global digital distribution, AI is able to amplify harms worldwide. What happens when systems trained for one geographical context are deployed in another without proper considerations and adjustments? Where has the AI system in question been developed? Does it take into account local data on the population with which it is going to interact?
Trustworthiness	How robust, accurate and efficient is the AI? Have impact assessments been carried out? Is there public information about the AI's accuracy? Does the AI use third-party software created by a known or trustworthy developer? Example: In 2018, an American Civil Liberties Union (ACLU) test of Amazon's Rekognition software used a default 80 per cent threshold confidence metric for recognizing faces. The ACLU claimed that, with this threshold applied in its test, the software had mistaken 28 members of Congress for other people who had been arrested on suspicion of having committed a crime.[153]
Obfuscation	How opaque are the AI, its processes and/or results to its creators, users, operators and/or recipients? Does the AI operate in a black box? This implies different facets: **Opacity:** There is often a lack of transparency in terms of accessing the inner workings of AIs – whether for reasons of complexity for those who lack technical knowledge in the area (*technical opacity*) or due to legal provisions that limit access. This can transform AIs into 'legal black boxes' (*legal opacity*). **Inscrutability:** Even if a person has the technical knowledge to understand AI at a deep level, neural networks are so massive and complex that it may be impossible to understand how a system arrives at a certain output. **Explainability, interpretability and traceability:** This refers to how AI systems provide justification for their actions, and how their 'reasoning' can be recreated for testing.
Bias	Are there conscious or unconscious biases in the selection of the data, development of the model or the interpretation of the results? Many examples of bias have been quoted in this essay, but these are just a fraction of the known and unknown cases in which biased AI is causing harm across different sectors, both public and private.
Accountability	Can algorithms be audited (*auditability*)? How will the law transfer the consequences of damage caused by AI to those responsible ? How can individuals harmed by AI have access to meaningful, feasible and effective remedies (*liability*)?
Fairness, equity and inclusion	How do AI systems affect society? Do they attempt mathematically to portray a model of reality in a fair way (*fairness*)? Do they proactively attempt to correct existing inequities, so that these inequalities and their effects are not transferred to the digital sphere (*equity*)? Who is represented and who is missing in AI outputs (*inclusion*)?

153 Levin (2018), 'Amazon face recognition falsely matches 28 lawmakers with mugshots, ACLU says'; and Snow, J. (2018), 'Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots', ACLU NorCal, 26 July 2018, https://www.aclunc.org/blog/amazon-s-face-recognition-falsely-matched-28-members-congress-mugshots.

09
Common goals and cooperation – towards multi-stakeholderism in AI

Responsible development of AI cannot occur in silos. It needs to be jointly and cooperatively guided, through global processes for reconciling competing interests and agreeing priorities. Now is a critical time for action, while innovations such as generative AI are still in their infancy.

Mira Lane and Stacey King

In late December 2023, a group of MIT researchers published their discovery of a new class of drug compounds that could kill antibiotic-resistant MRSA, a deadly form of drug-resistant staph bacteria. To accomplish this, the researchers used artificial intelligence (AI) to aid in their discovery and calculate potency predictions, an approach that also opens the door to designing more useful drugs in the future.[154] Drug discovery is just one of the many frontiers where experts expect AI to change current paradigms: not only in science but also in work, communication, media and the knowledge economy. A new wave of powerful technologies is showcasing just how far AI has come, both in interpreting almost unimaginably complex data and – in some applications – emulating human-like thought processes.[155]

AI-fuelled change evokes a spectrum of emotions. Leaps forward in medicine and science bring enormous excitement; threats of disruption and questions of safety bring apprehension and concern.[156] These mixed emotions are decades old: fears of technological disruption have run in parallel to the growing centrality of AI to our daily lives.

154 Wong, F. et al. (2023), 'Discovery of a structural class of antibiotics with explainable deep learning', *Nature* 626, pp. 177–85, 20 December 2023, https://doi.org/10.1038/s41586-023-06887-8.
155 Hagendorff, T., Fabi, S. and Kosinski, M. (2023), 'Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT', *Nature Computational Science* 3, pp. 833–38, 5 October 2023, https://doi.org/10.1038/s43588-023-00527-x.
156 Sartori, L. and Bocca, G. (2022), 'Minding the gap(s): public perceptions of AI and socio-technical imaginaries', *AI & Society*, Volume 38, pp. 443–58 (2023), 26 March 2022, https://doi.org/10.1007/s00146-022-01422-1.

Empowerment and disruption

Just like any revolutionary general-purpose technology, AI will have diverse impacts. In part, it will empower; in part, it will disrupt and present dilemmas.

In terms of empowerment, AI can make resources and skills available far more widely. For example, AI in translation services has bridged communication gaps on a global scale,[157] fostering collaboration across diverse cultures. 'Generative AI' – a type of artificial intelligence that can generate original content, ranging from text, images and music to code and synthetic data, after learning from a set of data inputs – presents an opportunity for more people to draw on legal, educational or medical expertise that would previously have been unaffordable or inaccessible. Generative AI and new learning tools are also revolutionizing teaching.[158] The advancements offer personalized and adaptive study experiences, catering to individuals' learning styles and pace, redefining the educational landscape, and challenging conventional structures and norms of knowledge acquisition and dissemination. For example, AI can tailor educational content to meet the individual needs of students, adjusting the difficulty level, suggesting resources based on learning styles, and providing personalized feedback to help students (and teachers) improve. AI's roles in medical research, such as AlphaFold's contribution to predicting the structures of proteins with remarkable accuracy,[159] and in the democratization of coding skills through intelligent coding assistance[160] demonstrate AI's capacity to lower the barriers to entry in many spheres.[161]

If AI is poised to drive a revolution in what is possible in science and technology, it is equally poised to disrupt. Economies, organizational structures, social contracts, and individual beliefs and opinions are all set to change as the next generation of AI becomes widespread. These changes will bring a responsibility to manage the risks and challenges posed by AI: ranging from the potential for AI-generated content to deviate from factual accuracy (leading to what are termed 'hallucinations', or misrepresentations of reality[162]) to the redefinition of jobs and conventional instructional roles and approaches. The impact of AI on labour markets – where, for example, its efficiency and automation capabilities can lead to significant shifts in employment patterns – necessitates a re-evaluation of job roles and skill requirements.[163]

---

157 Doherty, S. (2016), 'The Impact of Translation Technologies on the Process and Product of Translation', *International Journal of Communication*, Vol. 10, pp. 947–69, https://ijoc.org/index.php/ijoc/article/view/3499/1573.
158 Lim, W. M. et al. (2023), 'Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators', *The International Journal of Management Education*, 20(2), July 2023, https://doi.org/10.1016/j.ijme.2023.100790.
159 Hassabis, D. (2022), 'AlphaFold reveals the structure of the protein universe', Google DeepMind blog, 28 July 2022, https://deepmind.google/discover/blog/alphafold-reveals-the-structure-of-the-protein-universe.
160 Sundberg, L. and Holmström, J. (2023), 'Democratizing artificial intelligence: How no-code AI can leverage machine learning operations', *Business Horizons*, 66(6), pp. 777–88, https://doi.org/10.1016/j.bushor.2023.04.003.
161 Kanbach, D. K. et al. (2023), 'The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective', *Review of Managerial Science*, 13 September 2023, https://doi.org/10.1007/s11846-023-00696-z.
162 See, for example, Sharun, K. et al. (2023), 'ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references – a preliminary study', *Annals of Medicine & Surgery*, 85(10), pp. 5275–78, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10553015.
163 Frey, C. B. and Osborne, M. A. (2017), 'The future of employment: How susceptible are jobs to computerisation?', *Technological Forecasting and Social Change*, 114, pp. 254–80, https://doi.org/10.1016/j.techfore.2016.08.019.

Now is a critical time, while this next stage in the technology is still in its relative infancy, for governments, regulators, businesses, academia and the public to educate themselves and one another about this technology and its impact, and together to prepare for and negotiate the changes – positive and negative – AI will bring. Critical to this will be managing the pressures and competing goals that could impede a coordinated and coherent response, whether across industry or at national or international level.

Rising tensions

The heightened focus garnered by the very public commercialization of generative AI tools means businesses face a more competitive landscape. There is an increased emphasis on speed to market, as companies strive to gain a commercial advantage by adopting more powerful AI tools. Embracing AI can provide businesses with enhanced infrastructure. It can facilitate advancements in products or processes, help to attract top talent and employees, expand user or customer bases, and lead to valuable insights and possibilities. However, the growth of AI also threatens to affect trust between businesses, potentially weakening prospects for cooperation critical to effective multi-stakeholder processes.

AI is also poised to disrupt relations between governments. Pressure to ensure that the economic, market and national security benefits of these technologies are reaped locally potentially places governments in a race against one another. Developing regulations that reward national or regional AI development – while placing constraints on the import or export of AI technology – will heighten competition between nations. It could also hinder trust between business and government, for example encouraging more fractured and protectionist policies if governments – suspicious of the reach and intentions of transnational tech firms – seek to restrain such firms' borderless operations.

AI will also ask new questions of the relationship between citizens and states. Throughout history, shifts in technology have resulted in disruptions and economic hardships for individuals. Governments have often been forced to adapt accordingly, to ensure continued provision of citizens' basic needs in relation to safety, prosperity and economic opportunity. As AI alters the social contract in new ways, in fields from employment to politics to security, governments will again have to be responsive. And they will have to manage this disruption while also confronting the new risks posed by states that do *not* share a common purpose – think Russia, for example – for which AI potentially offers a tool to strengthen their power, sow division or enable new forms of international aggression.

Where the rise of AI differs most from previous technological shifts, however, lies in the *nature* of AI itself. AI is inherently complex, is evolving rapidly, and for the first time seeks to mechanize the human 'thought' process. This can make it difficult to fully understand or explain. It also makes the trajectory of AI development hard to predict, complicating policy decision-making concerning its risks and impacts, and creating new, existential fears among individuals. The future landscape is unknown, and the role that cognitive labour may have in an AI-driven world

is uncertain. With forecasts ranging from the extremes of machine dominance to an AI-powered utopia (with the likely reality being an unknown state somewhere between the two), we are in some ways navigating uncharted waters.

As multiple pressures and competing interests build around the development of AI, it will be critical for humanity to find a common path and pursue the collective interest. In many ways, AI is only as good as the training it is given, the rules and regulatory frameworks that govern its operations, and the specific applications in which it is utilized. Reflecting this breadth in its collective governance will be crucial.

Towards cooperation

There are several immediate steps we can take in navigating the competing pressures around AI development, and in directing that development constructively and towards a common goal.

First, governments and industry should make use of existing capacities. Existing laws on privacy, intellectual property, discrimination, competition and transparency all touch on questions of AI development and deployment. There are skills and expertise found in international treaty organizations, multilateral institutions, standards bodies, research consortiums and open-source communities that can support global cooperation. Existing principles on responsible innovation, and frameworks used by businesses and non-governmental organizations (NGOs), could serve as cross-industry models for businesses that both build and use AI or incorporate it in their operations.

Second, where waters are truly uncharted, it is imperative that stakeholders cooperate to identify and address genuine gaps or deficiencies within existing regulatory frameworks, standards and self-governance models. Partnerships between technical standards bodies and regulators could provide greater understanding of whether desired regulations can be put into realistic and executable practices. Such partnerships could support innovation while providing much-needed clarity to enable businesses of all sizes to comply with expectations and best practices for safe and responsible AI development.

Third, cooperation turns on equal access and transparency. This means ensuring wider availability of adequate physical 'compute' resources, shared public datasets and AI expertise (access to information and training that enable the development of AI), so that a global community of academic researchers, open-source communities and NGOs can contribute to AI development in an environment in which research and policy formation occur as transparently as possible. Governments that seek to encourage development of AI locally can partner with each other to provide repositories of public data, 'sandbox environments' in which to test models safely, and forums in which to discuss responsible AI principles. Companies that develop and use these technologies will require (a) alignment on quality, safety, reliability and fairness benchmarks; (b) the ability to publicly share details around training data without putting their intellectual property or users' privacy at risk; and (c) neutral forums for developing 'watermarks' and other mechanisms for indicating the types of content users may interact with. These

steps will support the development of AI that better protects societies from AI misuse or AI-driven misinformation, while driving greater understanding of the benefits of AI-generated content and outputs.

The message is clear. The work to develop AI cannot be done in silos, which means we need to overcome the competing pressures that drive 'silo-ization'. Leaving critical decisions in this area to be made solely by those who develop the technology – in effect, trusting that answers to complex problems will be solved later – is not an option. Nor can we govern AI without understanding it. The debate needs to expand beyond those small sections of society that traditionally develop digital technology or regulations. It needs to include voices that offer a more diverse representation of society – not only across age, gender and race, but also in terms of geography, profession, culture and economic status.

Crucial to this is recognizing that AI is no longer just a tool, but a general-purpose technology requiring collective governance. This means finding common ground through research, regulation and international cooperation, and agreeing on global priorities while the latest generative AI technologies are still relatively nascent.

All this may sound overwhelming, even insurmountable. But it is not. Humanity has worked through the impacts of complex technologies before: the introduction of the printing press, electricity, the railways and the internet. For over a decade, we have lived with AI being integrated in our lives in ever-increasing ways. We have already started building part of the social contract needed to govern AI. We are not starting from scratch. By committing to a common purpose and investing in the infrastructure of cooperation, we have the potential to shape a more positive and flourishing future society in which AI is used to the benefit of all.

About the authors

Alex Krasodomski leads the Chatham House Digital Society Initiative's research on digital public infrastructure, state and private sector cooperation and competition on technology provision, and public AI. He is a fellow at the Institute for Strategic Dialogue (ISD), Demos and the Public AI Network, and is a co-organizer of AI Palace.

Until June 2022, he was the research director at Demos, and director of the Centre for the Analysis of Social Media, during which time he authored more than a dozen major reports on digital election integrity, content moderation practices, digital regulation and the intersection between tech and politics.

Arthur Gwagwa is a research fellow in the Ethics of Socially Disruptive Technologies programme, based at Utrecht University. He has varied research interests in AI governance, including regulation, geopolitics, philosophical ethics, and intercultural non-domination perspectives that draw from the perspectives of under-represented groups, including indigenous groups of Africa, North America and the Antipodes. He is also a member of the Chatham House Digital Society Initiative's Responsible AI Taskforce, a board member of the Stanford *Journal of Online Trust and Safety*, an expert on Zimbabwe for UNESCO's Recommendation on the Ethics of Artificial Intelligence, and a member of the AI Academic Network of the Center for AI and Digital Policy. Arthur is a globally recognized human rights lawyer with a particular interest in human rights in the digital age; he is acknowledged as one of Africa's pioneers in that area of law and as a change-maker in digital rights. In his spare time, Arthur is a college admissions counsellor.

Brandon Jackson is an expert in product innovation at the Public AI Network, an international coalition of researchers working to make the case for public investment into AI. He has over 15 years of experience building public-interest technologies in mission-driven start-ups in both the US and UK tech ecosystems. He has a BA in computer science from Yale University and an MPhil in the history of science and technology from the University of Cambridge, where his research centred on the public adoption of new technologies such as radios.

Elliot Jones is a senior researcher at the Ada Lovelace Institute. He is currently working on approaches to foundation model evaluation and foundation model use in the public sector. At Ada, he has led research on the ethics of public service media recommendation systems, vaccine passports and digital contact tracing apps. He has previously been a researcher at Demos's Centre for the Analysis of Social Media, and a summer fellow at the Centre for Governance of AI.

Stacey King is a director of trust strategy at Google, where she is responsible for developing and leading strategies that uphold Google's values in response to emerging regulations. Prior to joining Google, Stacey served as both the Alexa Trust policy principal – developing principles, guidelines and technical solutions for the ethical development of AI, data and content – and as the business and technical leader of an Amazon subsidiary/incubation group. Stacey is a peer reviewer for the journal *AI and Ethics*. She recently completed a three-year visiting policy fellowship at the Oxford Internet Institute, researching notions of authorship, creativity, ownership and the public domain in relation to AI-generated works. She has a background in strategy, history and law.

Mira Lane is the senior director and founder of the Envisioning Studio at Google. Mira runs a multidisciplinary team focusing on showcasing the inspiring potential for advanced technologies to benefit people and societies. Prior to joining Google, Mira was the partner director and founder of Ethics & Society at Microsoft. There, her team was responsible for guiding technical and experience innovation towards ethical, responsible and sustainable outcomes. Mira holds numerous patents across platforms and collaborative interfaces. She has a background in art, computer science and mathematics. Her art has been featured in film festivals and galleries.

Micaela Mantegna is a video game lawyer and activist who is internationally recognized for her expertise in AI ethics, extended reality (XR) policy, and the complex relationship between AI, creativity and copyright law. She is an affiliate at the Berkman Klein Center at Harvard University, and serves on the World Economic Forum's Metaverse Council, Chatham House's Responsible AI Taskforce, and the Scientific Committee of UAMetaverse Chair. She earned fellowships at Google Policy in 2017 and TED in 2022, alongside Salzburg Global and Datasphere Initiative fellowships, and was the lead drafter of the ethics chapter of Argentina's National AI Plan in 2019. Currently, she lends her policy expertise to organizations and governments, contributing to the development of metaverse policy around the world.

Thomas Schneider is ambassador and director of international affairs at the Swiss Federal Office of Communications (OFCOM) in the Federal Department of the Environment, Transport, Energy and Communications (DETEC). He is a long-standing expert in digital governance. For the last 20 years, since the UN World Summit on the Information Society in 2003, he has been leading the Swiss delegation on internet and digital governance issues in various international forums. He is the chair of the Council of Europe's Committee on Artificial Intelligence, which in 2022 was mandated to negotiate a legally binding instrument on AI. From 2014 to 2017, he was the chair of the Governmental Advisory Committee of the Internet Corporation for Assigned Names and Numbers (ICANN), and in this role negotiated the compromise among governments regarding the 'IANA Stewardship transition', the biggest reform in the ICANN system. He was responsible for the organization of the 12th UN Internet Governance Forum (IGF) in Geneva in December 2017 on behalf of the Swiss government, and was co-chair of the IGF's Multistakeholder Advisory Group in 2017. From 2020 to 2022, he was vice-chair of the OECD's Committee for Digital Economy Policy.

Kathleen Siminyu is an AI researcher focused on natural language processing (NLP) for African languages. She works at the Mozilla Foundation as a machine learning fellow to support the development of a Kiswahili speech recognition dataset, and to build transcription models for end-use cases in the agricultural and financial domains. In this role, she is keen to ensure that the diversity of Kiswahili speakers – in terms of age, gender, accent and language variant/dialect – is catered for in the dataset and models created. She would welcome opportunities exploring the application of speech technologies in education.

Before joining Mozilla, Kathleen was regional coordinator of AI4D Africa, where she worked with machine learning and AI communities in Africa to run research programmes. One of these, a fellowship for African-language dataset creation, led to the creation of multiple African-language datasets. For this work, Kathleen was

listed as one of the *MIT Technology Review* 35 Innovators aged under 35 for 2022. She has extensive experience as a community organizer, having co-organized the Nairobi Women in Machine Learning and Data Science community for three years. She continues to organize as part of the committees of the Deep Learning Indaba and the Masakhane Research Foundation.

Alek Tarkowski is the director of strategy at Open Future, a European non-profit organization that works to advance public policies and civil society strategies around openness and protection of the digital commons. He has over 15 years of experience with public-interest advocacy, movement-building and research into the intersection of society, culture and digital technologies. He is a sociologist by training and holds a PhD in sociology from the Polish Academy of Science.

In 2010 he established Centrum Cyfrowe, one of the leading Polish organizations promoting openness and internet users' rights. Before that, he was a strategic adviser to the prime minister of Poland, co-authoring the 2009 report *Poland 2030* and the Polish official long-term strategy for growth.

In 2005, he co-founded Creative Commons Poland; since then he has been an active member of the Creative Commons network. He is currently a member of the board of directors of the Creative Commons organization.

Acknowledgments

The editor appreciates the valuable work of all the contributors to this paper and thanks them for their essays. An essay collection is a significant undertaking, and the authors' insight, warmth and patience have been exemplary.

At Chatham House, enormous thanks to Rowan Wilkinson, Marjorie Buchser and Isabella Wilkinson on the Digital Society Initiative team for their support and expertise, to Bronwen Maddox for her excellent foreword and encouragement, to Alex Vines for his advice, and to the wider group of international experts whose efforts and input improved the paper significantly. I am grateful to the Chatham House Responsible AI Taskforce, Carl Miller, Ellen Judson, Joshua Tan and the Public AI Network.

We would also like to thank the anonymous peer reviewers for their time and valuable contributions. This publication has been made possible by Google Search's generous support for global technology work at Chatham House, and I am particularly grateful to Erica Fitzpatrick for her ongoing trust and energy.

Finally, thank you to Jake Statham for his careful and considered editing, without which the collection would not have come together.

Cover image: Visitors watch a projection of Turkish-American artist Refik Anadol's AI-generated work at the Serpentine North Gallery in London, February 2024.
Photo credit: Copyright © Dan Kitwood/Getty Images

ISBN 978 1 78413 608 6
DOI doi.org/10.55317/9781784136086

Cite this paper: Krasodomski, A. (ed.) et al. (2024), *Artificial intelligence and the challenge for global governance: Nine essays on achieving responsible AI*, Research Paper, London: Royal Institute of International Affairs, https://doi.org/10.55317/9781784136086.

This publication is printed on FSC-certified paper.
designbysoapbox.com

Independent thinking since 1920

CHATHAM HOUSE

The Royal Institute of International Affairs
Chatham House
10 St James's Square, London SW1Y 4LE
T +44 (0)20 7957 5700
contact@chathamhouse.org | chathamhouse.org

Charity Registration Number: 208223