Research Paper

Digital Society Initiative

April 2023

Recalibrating assumptions on AI

Towards an evidence-based and inclusive AI policy discourse

Arthur Holland Michel

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

Summary

- Artificial intelligence (AI) policy is underpinned by a range of common assumptions about how AI will contribute to economic, military and societal advantage, how such 'AI power' can be harnessed, and how the technology's known risks can be averted. Many of these assumptions have become entrenched despite the fact that they do not represent the interests of all stakeholders and they misrepresent a growing body of evidence. Policies built upon these assumptions could lead to elevated risks for certain demographic groups, among other negative outcomes.
- As AI policies begin to take the form of hard rules and regulations, the common assumptions that underpin them must be able to accommodate new facts and be comprehensive enough to account for all possible risks. Most importantly, they must be representative of the interests of all stakeholders affected by the technology and the norms that govern it.
- Examples of dominant AI assumptions include the claim that AI is 'intelligent', that 'more data' is a requisite for better AI, that AI development is 'a race' among states, and that AI itself can be 'ethical'.
- These types of assumption are unyielding to contrary evidence, counterfactuals and nuance that challenges their standing. They extol the virtues of AI but neglect to consider its failures, while tending to serve a narrow but powerful set of interests. And they minimize perspectives from beyond the Global North and the male-dominated tech sector, as well as perspectives centred around non-technical approaches to societal problems.
- The dominance of these AI assumptions is problematic. The greater the gap between a policy assumption and the facts and people it supposedly represents, the greater the risk that the policy could result in harm – and the harder it is to draw from the full spectrum of options necessary to build robust and equitable regulations.
- None of this implies that policy assumptions are, in themselves, bad for AI policy discourse. But the most transformative and equitable AI policies will be those that engage actively with all uncomfortable counterpoints and with all under-represented perspectives – not just with the most prominent and powerful stakeholders.

- To offset the potentially harmful effects of unchallenged assumptions about AI, stakeholders should:
  - Recognize when an assumption that lacks firm, unequivocal evidence is being used as the basis for a policy, and provide a framework to consider that assumption's consequences and counterpoint(s);
  - Identify who these assumptions serve and consider whether those groups or individuals are representative of all stakeholders;
  - Explore alternative or additional policymaking assumptions; and
  - Establish a practice of pre-development risk assessments for proposed AI systems.

01 Introduction

The core assumptions of AI policy are in dire need of recalibration. Drawing from a wider range of evidence and perspectives would result in safe and more equitable outcomes for AI policy.

With startling consistency, artificial intelligence (AI) policy is underpinned by common assumptions about how AI will contribute to economic, societal and military advantage, how such 'AI power' can be harnessed, and how the technology's known risks can be averted. As AI policies across the globe pass from theory to practice in the coming years, these common assumptions must keep pace with the facts. They must also be clear-eyed enough to account for all possible risks. Most importantly, these assumptions must be representative of the interests of all stakeholders who will be impacted by the technology and the rules that govern it.

So far, this has not been the case. Common AI assumptions are so often repeated that one would think they reflect a preponderance of evidence. But many of these assumptions are more like opinion than truth. They are dominant simply because they are unyielding to facts that challenge their status, hostile to caveats and inimical to nuance. And these views are not universal. Rather, they tend to reflect and serve a narrow but powerful set of interests, while minimizing perspectives from beyond the Global North and the male-majority tech sector, as well as perspectives centred around equality, sustainability and humanistic – in other words, non-technical – approaches to societal problems.

This does not bode well for the coming years of AI policy. The greater the gap between a policy assumption and the facts and people it supposedly represents, the greater the risk that measures built upon that assumption could result in harm. The more tightly that a policy adheres to assumptions that serve the narrow interests of one set of stakeholders, the less likely it is that the benefits of that policy will be distributed widely or fairly. The less amenable the policy sphere makes itself to voices that do not buy its received truths, the harder it will be to draw from the full spectrum of solutions that are needed to build robust and equitable outcomes.

This paper makes a bid to recalibrate the AI policy discourse. It highlights, analyses and offers counterpoints to four core assumptions of AI policy: 1) that AI is 'intelligent'; 2) that 'more data' is a requisite for better AI; 3) that AI development is 'a race' among states; and, 4) that AI itself can be 'ethical'. It focuses on these four assumptions because they have gone particularly unchallenged in policy

documentation, and because they demonstrate how real harms can result from policy that is built upon assumptions that negate counterpoint perspectives. In challenging these assumptions, the paper offers a rubric for addressing other problematic AI assumptions. By illustrating how a more evidence-based, inclusive discourse yields better policy, it advocates for an ecosystem of policy innovation that is more structurally diverse and intellectually accommodating.

Some disclaimers

Though this paper critiques some of the fundamental assumptions underpinning government efforts to get ahead in AI, it does not advocate against governments taking seriously the disruptive potential of these technologies. States and their citizens have a sovereign right – within the boundaries of national and international law – to reach for the opportunities of novel algorithmic systems. But while optimism and a competitive spirit may be key drivers of technical progress, they are a poor basis for safety-critical regulations. This paper therefore advocates for responsible policy that does not withhold from deeming certain applications of AI as undesirable, or certain institutions as not being 'AI ready'. It calls for parties to recognize that, given the power of AI (and the power of the organizations wishing to use it), the cost of acting with a critical eye may often be far lower than the cost of acting unquestioningly on an overly optimistic assumption that later turns out to be wrong.

Common AI assumptions are so often repeated that one would think they reflect a preponderance of evidence. But many of these assumptions are more like opinion than truth.

This is not to say that this paper proposes its own anti-risk dogma. It is possible that as a society we can accept a degree of risk in exchange for the possibility that someday AI might make good on its promise. But that is only an acceptable conclusion to reach if it has the buy-in of all relevant stakeholders – especially those who are most likely to suffer from the risks. If some parties to the debate object to the mere suggestion that a particular application of AI might not be worthwhile or that attaining truly 'ethical AI' is not necessarily a *fait accompli*, it will be impossible to achieve such consensus.

Nor does this paper argue that all policy assumptions are, in and of themselves, a bad thing. Any policymaking for a nascent technology will rely on some degree of supposition about what that technology will and will not do in the future. Certain widely held AI assumptions have already proven to be useful. For example, it is often noted in AI policy documents that all AI systems are liable to exhibit biases against certain groups. While there are, of course, exceptions to this assumption, policymakers can use it to ensure that the possibility of bias is never neglected in proposed measures and instruments.

Yet even in this case, if those holding the assumption refuse to engage with counterpoints and emerging contrary evidence, it could become problematic to continue to adhere policy strictly to that assumption.

So, let this much be clear: *any* AI policy assumption is liable to become harmful dogma if not held open to honest, good-faith challenges. The most transformative AI policies will be those that engage with uncomfortable counterpoints to *all* predominant assumptions and engage with *all* under-represented perspectives – not just with the loudest voices in the room.

This paper is intended for a cross-cutting audience of parties to the discourse on AI strategy and policy, including policymakers, private sector stakeholders, commentators and advocacy groups. It is based on a review of national AI strategies, policy documents, AI bills,[1] technical literature and critical commentary. Input was also collected through a virtual expert roundtable that was hosted by Chatham House on 2 March 2022.

Four AI assumptions and their counterpoints

Assumption: Artificial intelligence has unlimited potential to execute any task that ordinarily requires human intelligence, input, oversight and judgment.

Counterpoint: The technologies currently referred to as 'artificial intelligence' are inherently limited in their capacity to replicate human intelligence. Rather, they have only demonstrated themselves to be capable of imitating narrow facets of human intelligence in certain narrow tasks, and they could continue to be ineffectual for certain applications for many years to come.

Assumption: A principal enabler of AI development and deployment is data. Therefore, states wishing to increase their AI capacity should endeavour to collect, consolidate and distribute the greatest volume of relevant data.

Counterpoint: Not all applications of AI will necessarily benefit from the collection, centralization and distribution of data. Furthermore, any data collection and distribution activity carries serious risks. In some cases, those risks may outweigh the anticipated gains the data might yield for AI development.

Assumption: In order to succeed in international power competition, states must develop and deploy AI more widely and more quickly than their adversaries and peers.

Counterpoint: A race-like approach to technology development could stand at odds with a state's capacity to adopt AI in a way that truly serves the common good.

Assumption: Ethical principles can be encoded into AI.

Counterpoint: Achieving 'ethical AI' requires expansive measures that extend far beyond strictly technical fixes, including – potentially – uncomfortable organizational and societal reform.

1 For a useful repository of these documents, see OECD (undated), 'National AI policies & strategies', https://oecd.ai/en/dashboards/overview.

02 Artificial ‘intelligence’ – a problematic definition

Artificial intelligence is not actually intelligent. Policy that treats AI as though it is intelligent could have significant ramifications and could result in significant material risks for individual citizens and companies alike.

Assumption: Artificial intelligence has unlimited potential to execute any task that ordinarily requires human intelligence, input, oversight and judgment.

Counterpoint: The technologies currently referred to as ‘artificial intelligence’ are inherently limited in their capacity to replicate human intelligence. Rather, they have only demonstrated themselves to be capable of imitating narrow facets of human intelligence in certain narrow tasks, and they could continue to be ineffectual for certain applications for many years to come.

In the policy sphere, AI is widely characterized as computerized technologies capable of executing tasks that ordinarily require human intelligence.[2] This represents an ideological, rather than technical, notion[3] that extends the scope

2 Portugal INCoDE (2019), ‘AI Portugal 2030: Portuguese National Initiative on Digital Skills’, https://www.incode2030.gov.pt/en/ai-portugal-2030; National Science & Technology Council (2019), *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, Washington, DC: Executive Office of the President of the United States, p. 1, https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf; United Kingdom Office for Artificial Intelligence (2022), ‘National AI Strategy’, https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version; Singapore Smart Nation and Digital Government Office (2019), ‘National AI Strategy’, https://www.smartnation.gov.sg/initiatives/artificial-intelligence; France’s strategy offers a somewhat more cautious definition, though it falls short of a rebuke of those offered by other strategies, see Villani, C. (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*, https://www.aiforhumanity.fr.

3 Lanier, J. and Weyl, E. G. (2020), ‘AI is an Ideology, Not a Technology’, *WIRED,* 15 March 2020, https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology.

of AI policy beyond the technology that exists today to include unlimited technologies that may not yet exist for years to come, if they ever do.[4] In some governments it is even a matter of policy to assume that the intelligence of computerized systems will continue to grow indefinitely, to the point of achieving 'artificial general intelligence' that matches or exceeds human mental capacity.[5] It is in light of this definition that AI is commonly likened to electricity:[6] a general-purpose technology of boundless potential,[7] which will have, to quote one strategy, 'a transformational impact on the whole economy'.[8] This common policy-level view of AI is technologically problematic, and it risks derailing AI policy in several key ways.

Natural limits

It is indisputable that algorithmic technologies have proven to be effective – and in some cases transformative – in certain applications such as digital marketing, social media, web search, some domains of finance and computer vision. Nevertheless, recent years have also provided ample evidence that AI still falls far short of being an artificial version of human intelligence. AI has exhibited persistent failures across a wide set of domains. In one way or another, these failures stem from the fact that these systems do not yet replicate the fundamental capacity of real human intelligence to account for ambiguity and anomalies,

---

4 As Michael Atleson, an attorney for the US Federal Trade Commission's Division of Advertising Practices puts it, artificial intelligence is 'an ambiguous term with many possible definitions… But one thing is for sure: it's a marketing term', see Atleson, M. (2023), 'Keep your AI claims in check', Federal Trade Commission Business Blog, 27 February 2023, https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.
5 From the UK's AI strategy: 'we take the firm stance that it is critical to… take seriously the possibility of AGI and "more general AI"', United Kingdom Office for Artificial Intelligence (2022), 'National AI Strategy,' p. 17; National Science & Technology Council (2019), *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, pp. 10–11; Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025', p. 92, https://cbddo.gov.tr/en/nais; President of the Russian Federation (2019), *On the Development of Artificial Intelligence in the Russian Federation* [translated by CSET Georgetown], Moscow: President of the Russian Federation, https://cset.georgetown.edu/wp-content/uploads/Decree-of-the-President-of-the-Russian-Federation-on-the-Development-of-Artificial-Intelligence-in-the-Russian-Federation-.pdf. From Finland's AI Strategy: 'In the future, we will see AI systems that are much more aware of their environment and are able to adapt to change. This will inevitably lead to systems that are much more like humans and are able to adjust to changes around them', and 'In the long term (in 20 to 50 years), AI may reach the performance level of humans or even exceed our capabilities in most tasks'. See Steering group and secretariat of the Artificial Intelligence Programme (2019), *Leading the way into the era of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019*, Helsinki: Ministry of Economic Affairs and Employment, p. 32 and p. 36, https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161688/41_19_Leading%20the%20way%20into%20the%20age%20of%20artificial%20intelligence.pdf.
6 Horowitz, M. (2018), 'Artificial Intelligence, International Competition, and the Balance of Power', Texas National Security Review 1(3), pp. 36–57, https://doi.org/10.15781/T2639KP49.
7 Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025'. Also see the 'AI+X' approach discussed in India's 2018 AI strategy discussion paper: Kumar, A., Shukla, P., Sharan, A. and Mahindru, T. (2018), *National Strategy for Artificial Intelligence #AIforAll*, New Dehli: NITI Aayog, p. 22, https://indiaai.gov.in/documents/pdf/NationalStrategy-for-AI-Discussion-Paper.pdf; Villani (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy;* Portugal INCoDE (2019), 'AI Portugal 2030: Portuguese National Initiative on Digital Skills'. A European Commission proposal states that AI 'can contribute to a wide array of economic and societal benefits across the entire spectrum of industries and social activities', see 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', https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. According to Singapore's national strategy, 'AI is a general purpose technology. It has applicability in any field', see Smart Nation and Digital Government Office (2019), 'National AI Strategy'.
8 United Kingdom Office for Artificial Intelligence (2022), 'National AI Strategy'.

to understand concepts and their relation to each other, to adapt to new information, to grasp the difference between truth and untruth, and to take non-numerical factors into account in problem-solving.

AI's most commonly cited breakthroughs in digital environments cannot be used as proof that AI will succeed in the real physical world, let alone in other domains, or in safety-critical settings.

AI's most commonly cited breakthroughs in digital environments – computers that beat humans at games like chess, go and Starcraft, for example, or that can design new proteins and medicines, or generate digital content – cannot be used as proof that AI will succeed in the real physical world, let alone in other domains, or in safety-critical settings.[9] In other realms, the research community has experienced endemic failures in achieving the same performance that machine learning exhibits in testing when it is deployed in real-world conditions.[10]

Proof of AI's shortcomings is plentiful. Large language models, which may be capable of writing a passable high school essay or engaging a user in a coherent conversation, consistently make highly unpredictable errors despite being trained on vast and extensive volumes of human language.[11] In transportation, despite hundreds of billions of dollars in investment, autonomous vehicles have yet to be deployed at scale, as they continue to be prone to failures when encountering specific unfamiliar conditions and situations on the road.[12] Uber, a presumed leader in the autonomous vehicle field, abandoned its self-driving vehicle division in 2020.[13] Free-ranging autonomous robotic systems with no human aboard remain confined mostly to experimental settings.

---

9 Sam Altman, the CEO of the AI company OpenAI and a champion of the theory that AI will eventually exceed human general intelligence, recently tweeted, Altman, S. (@sama) via Twitter (2022), 'ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. It's a mistake to be relying on it for anything important right now. it's a preview of progress; we have lots of work to do on robustness and truthfulness', 11 December 2022, https://twitter.com/sama/status/1601731295792414720.
10 Liao, T., Taori, R., Raji, D. and Schmidt, L. (2021), 'Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning', in Vanschoren, J. and Yeung, S. (eds) (2021), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html; Sohn, E. (2023), 'The reproducibility issues that haunt health-care AI', *Nature*, 9 January 2023, https://www.nature.com/articles/d41586-023-00023-2.
11 Ganguli, D. et al. (2022), 'Predictability and Surprise in Large Generative Models', https://arxiv.org/pdf/2202.07785.pdf.
12 Clarke, L. (2022), 'How self-driving cars got stuck in the slow lane', *The Guardian*, 27 March 2022, https://www.theguardian.com/technology/2022/mar/27/how-self-driving-cars-got-stuck-in-the-slow-lane; Todd, L. (2023), *Autonomous Vehicle Implementation Predictions Implications for Transport Planning*, Victoria: Victoria Transport Policy Institute, https://www.vtpi.org/avip.pdf.
13 Marshall, A. (2020), 'Uber Gives Up on the Self-Driving Dream', WIRED, 7 December 2020, https://www.wired.com/story/uber-gives-up-self-driving-dream.

In medicine, another highly touted critical application area, high-profile AI programs have so far yielded meagre performance gains compared to traditional systems or measures, and have also resulted in excess harms.[14] Several states have used the onset of the COVID-19 pandemic to call for quicker AI adoption, arguing that the technology could be useful in medical response and epidemiological forecasting,[15] though studies have shown that AI experiments related to COVID-19 have largely fallen short of expectations.[16]

To be sure, in any given area to which AI is being applied, it is possible to point to experimental efforts that have shown encouraging early performance. And it is undeniable that someday AI will succeed in areas where it currently struggles. But we do not have hard evidence as to which applications these will be. We also cannot say with certainty when these barriers to AI's success will fall. AI progress has never tracked along a linear growth curve – rather, it has moved along a series of largely unpredictable 'AI springs' and 'AI winters'. If we are still in the early days of a lengthy AI boom, these breakthroughs could be right around the corner. But if we are already witnessing the waning days of this AI boom cycle, they could still be a long way off.

Proven risks, conjectured rewards

Because the predominant understanding of AI assumes that significant advances are always imminent, there is a tendency in AI policy literature to grant the technology's expected benefits equivalent weight as its known shortcomings and risks. For example, one strategy notes that 'algorithmic risk assessment tools… have the potential to improve consistency and predictability' in pre-trial detention, sentencing and bail decisions while noting that 'the use of AI within the justice sector also has considerable implications for ethics, human rights and the rule of law'.[17] At face value, this would appear to be a balanced characterization. However, the strategy misrepresents the evidence: there is ample proof of the technology's harms and

---

14 Konam, S. (2022), 'Where did IBM go wrong with Watson Health?', QZ, 2 March 2022, https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health; Simonite, T. (2021), 'An Algorithm That Predicts Deadly Infections Is Often Flawed', WIRED, 21 June 2021, https://www.wired.com/story/algorithm-predicts-deadly-infections-often-flawed; Sohn (2023), 'The reproducibility issues that haunt health-care AI'. Researchers have also found widespread flaws in the documentation provided for medical AI systems, which undermines efforts to rate these on criteria such as fairness, transparency and robustness across different demographic patient groups. Lu, J. H. et al. (2021), 'Low adherence to existing model reporting guidelines by commonly used clinical prediction models', medRxiv, 23 July 2021, https://www.medrxiv.org/content/10.1101/2021.07.21.21260282v1.

15 United Kingdom Office for Artificial Intelligence (2022), 'National AI Strategy'; Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025'; 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'; Government of Ireland (2021), *AI – Here for Good: A National Artificial Intelligence Strategy for Ireland, Dublin: Department of Enterprise*, Trade and Employment, p. 44, https://enterprise.gov.ie/en/Publications/Publication-files/National-AI-Strategy.pdf; Van Roy, V., Rosetti, F., Perset, K. and Galindo-Romero, L. (2021), 'AI Watch – National strategies on Artificial Intelligence: A European perspective', 2021 edition, Luxemburg: Publications Office of the European Union, pp. 17–18, doi:10.2760/069178.

16 Roberts, M. et al. (2021), 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans', *Nature Machine Intelligence* 3, pp. 199–217, https://doi.org/10.1038/s42256-021-00307-0; Wynants, L. et al. (2020), 'Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal', *BMJ*, 2020; 369, https://doi.org/10.1136/bmj.m1328; Nixon, K. et al. (2022), 'Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation', *The Lancet Digital Health*, 4(10), E699-E701, 1 October 2022, https://doi.org/10.1016/S2589-7500(22)00167-4.

17 Government of Ireland (2021), *AI – Here for Good: A National Artificial Intelligence Strategy for Ireland*, p. 44.

fairly uneven evidence of proven benefits.[18] Achieving a true balance between the actual benefits and risks of predictive policing would likely depend on technological breakthroughs that remain speculative.

The risks and benefits of AI also tend to track along vastly different scales. An improvement in efficiency that may come from the use of a fraud detection algorithm is in no way comparable to the harm that results when that system leads to a wrongful accusation. The gains that an HR department might see from using a hiring algorithm are moot if that algorithm systematically (and from the user's perspective invisibly) privileges applicants from a particular demographic group. Treating risk as something that can be weighed against benefit is therefore misleading. At a practical level, it potentially undermines the capacity for state mechanisms that sponsor or regulate AI development to differentiate AI tools that should be pursued from those that absolutely should not.[19]

In the absence of such discernment, the harms of misapplied AI are rarely distributed evenly. Because AI limitations often manifest themselves in ways that reflect bias, the use of AI in tasks for which it is technically ill-suited or in contexts lacking sufficient regulatory guardrails poses an elevated risk of harm to members of vulnerable or historically disadvantaged groups.[20] In addition, yet-to-be-proven AI technologies are more likely to be used against populations with less policy leverage to advocate for protections or moratoriums (consider, for example, early experiments involving AI for welfare fraud detection,[21] predictive policing[22] and surveillance in public housing[23]).

To engender the safest possible regulations, it would be more appropriate for generalizations about a technology's readiness to weigh AI's failures more heavily than its successes. This reframing will be especially helpful when these failures fall along consistent patterns that might suggest that its 'growing pains' are actually inherent limitations. Entities evaluating a proposed AI role might use the following

---

18 Wisser, L. (2019), 'Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing', *American Criminal Law Review*, 56(4), pp. 1811–832, https://www.law.georgetown.edu/american-criminal-law-review/in-print/volume-56-number-4-fall-2019/pandoras-algorithmic-black-box-the-challenges-of-using-algorithmic-risk-assessments-in-sentencing; Metz, C. and Satariano, A. (2020), 'An Algorithm That Grants Freedom, or Takes It Away', *New York Times*, 6 February 2020, https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html; Algorithm Watch (2020), 'Automating Society Report 2020', https://automatingsociety.algorithmwatch.org.

19 'Emotion recognition' technology, which is gaining widespread interest for use in areas such as surveillance, marketing and hiring, is potentially one such type of system. There is ample evidence to suggest that emotion recognition technology does not work and furthermore is founded on racist physiognomic theories that have been debunked by numerous studies across scientific disciplines. See Crawford, K. (2021), 'Artificial Intelligence is Misreading Human Emotion', *The Atlantic*, 27 April 2021, https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696. Despite these controversies, India's AI strategy, for example, lists emotion recognition as a near-term application of AI. See Kumar et al. (2018), *National Strategy for Artificial Intelligence #AIforAll*, p. 15.

20 For example, even the accelerated adoption of a fairly rudimentary algorithm for disbursing high school standardized testing grades in the UK in the summer of 2020 resulted in harms that disproportionately affected minority and lower-income populations. See Ofqual (2020), 'Awarding GCSE, AS & A levels in summer 2020: interim report', https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

21 Henley, J. and Booth, R. (2020), 'Welfare surveillance system violates human rights, Dutch court rules', *Guardian*, 5 February 2020, https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules.

22 Sankin, A., Mehrotra, D., Mattu, S. and Gilbertson, A. (2021), 'Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them', The Markup, 2 December 2021, https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them.

23 Fadulu, L. (2019), 'Facial Recognition Technology in Public Housing Prompts Backlash', *New York Times*, 24 September, https://www.nytimes.com/2019/09/24/us/politics/facial-recognition-technology-housing.html.

formula to judge whether it is worth exploring further: If it can be accomplished *without* increasing harms, to the best of one's knowledge, this application of AI has the potential to yield gains – however, if there is evidence that the technology will increase risks to some groups, further study would be necessary *prior* to the actual use or real-world testing of the technology.

In this spirit, some AI policies have proposed or enacted moratoriums on particular applications of AI, such as live biometric surveillance,[24] given that (according to these policies) no conjectured benefits yielded by such applications could outweigh their amply demonstrated risks for the foreseeable future. At a minimum, policies could be transparent about their acceptance of a certain degree of risk in the pursuit of a particular application.[25] In such cases, they should be specific about who will most likely be affected by these risks, and then seek the buy-in of those communities before proceeding.

Ethical unintelligence

The 'intelligence' framing of AI muddles the discourse on how to prevent and respond to AI harms. If one assumes that a computer could replicate human intelligence,[26] this further assumes that the system, like a human, could include ethical mores in the parameters that guide its actions. It is certainly true that in some cases, computation has fruitfully displaced human judgment (consider, for example, autopilots on aeroplanes). However, those systems generally only mimic a very narrow column of human intelligence related to information retrieval, rule-based processing and pattern recognition – none of which can be mapped to ethical reasoning. Meanwhile, when a sophisticated AI fails it often does so in ways that no human possessing a modicum of genuine intelligence ever would. This makes it hard to rate and account for AI's reliability using the same metrics and tools that existing regulatory frameworks use for humans and predictable systems, such as mechanical components, that are not based on probabilistic AI. A more grounded strategy to implement 'ethical AI' would acknowledge that AI systems themselves can never, for instance, be 'held accountable' or be 'trustworthy' in the human-ethics sense.

---

24 Reuters (2022), 'Italy outlaws facial recognition tech, except to fight crime', 14 November 2022, https://www.reuters.com/technology/italy-outlaws-facial-recognition-tech-except-fight-crime-2022-11-14; Miller, M. (2020), 'Democratic lawmakers introduce legislation banning government use of facial recognition technologies', The Hill, 25 June 2020, https://thehill.com/policy/technology/504583-democratic-lawmakers-introduce-legislation-banning-government-use-of-facial.

25 The European Commission proposes a classification system that distinguishes AI systems by their level of risk, with different requirements applied for each category; however, this rating system lacks granularity, and will apply only to rules governing *deployed* systems rather than rules for guiding investment and development strategy. See 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'.

26 Consider, for example, the Australia AI Action Plan, which states that 'AI is more than just the mathematical algorithms that enable a computer to learn from text, images or sounds. It is the ability for a computational system to sense its environment, learn, predict and take independent action to control virtual or physical infrastructure'. See Australian Government Department of Industry, Science, Energy and Resources (2021), 'Australia's AI Action Plan', p. 4, https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-action-plan.

The predominant understanding of AI as 'intelligent' also assumes that present-day challenges for ethical AI, such as a lack of system predictability and transparency, can be engineered away by increasing the intelligence of these systems.[27] In truth, increasing system 'intelligence' (which can increase the system's autonomy and open-endedness)[28] may in fact be more likely to compound ethical challenges. Large language models demonstrate this phenomenon; the popular AI-based chatbot ChatGPT, for example, has raised myriad questions relating to intellectual property and fair-use, how to algorithmically balance safety against free speech, and where to assign responsibility for harms. An alternative strategy for ethical AI might therefore focus on non-technical (or even non-AI-related) measures to combat harms: for example, anti-corruption measures to improve institutional accountability or racial justice campaigns to improve societal equity (see Chapter 5).

Dollars and sense

The assumption that AI will be as transformative and essential to every aspect of modern life as electricity can stifle any measure that would place limits on AI use or capacity, even if those measures have other direct benefits. For example, a number of national AI strategies are explicit that their intent is not so much to explore the possibilities and limitations of AI; rather it is, as one strategy put it, to 'drive AI adoption in the private and public sectors'[29] and, to quote another, to 'support the diffusion of AI across the whole economy'.[30] One strategy even proposes 'adding the requirement for AI-based solutions in the specifications of other strategic investments… financed from public funds'.[31] That is, the state's critical decisions on strategic investments will be informed, in part, on whether the proposed investment involves 'AI'. The implication being that if the state is considering two proposed investments, one which involves AI and another which does not, the former will receive priority.[32]

---

27 Consider, for example, the many headlines about AI systems that can 'explain themselves' to their human users.
28 Ganguli et al. (2022), 'Predictability and Surprise in Large Generative Models'.
29 Government of Ireland (2021), *AI - Here for Good: A National Artificial Intelligence Strategy for Ireland*', p. 7; Australian Government Department of Industry, Science, Energy and Resources (2021), 'Australia's AI Action Plan'; Steering group and secretariat of the Artificial Intelligence Programme (2019), *Leading the way into the era of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019*, p. 51.
30 United Kingdom Office for Artificial Intelligence (2022), 'National AI Strategy'.
31 Polish Ministry of Digital Affairs (2020), 'Policy for AI Development in Poland from 2020', https://oecd.ai/en/wonk/documents/poland-policy-for-ai-development-in-poland-from-2020-2020. For a broader discussion on European national investments in AI, see Van Roy et al. (2021), 'AI Watch – National strategies on Artificial Intelligence: A European perspective'.
32 As Cave et al. write, 'Hype bubbles can lead to disproportionate amounts of research funding being directed into a field because it is prominent in certain narratives, at the expense of other fields of research'. See Cave, S. et al. (2018), 'Portrayals and perceptions of AI and why they matter', London: Royal Society, p. 15, https://www.repository.cam.ac.uk/handle/1810/287193.

Putting aside the ethical perils of this mindset, it could also be economically risky. Citing forecasts that estimate trillions of dollars of positive economic impact stemming from AI,[33] government investments in AI research and development have mushroomed in recent years.[34] Yet studies have shown that while adoption of AI across industries is growing, a significant proportion of enterprises that have embarked on AI projects have yet to see any substantial gains from the technology.[35] Recent polling also indicates that a large share of the machine-learning models that are developed fail to reach deployment.[36] To be sure, any government effort to make technological breakthroughs requires some tolerance for financial risk. But as the current AI boom cycle enters its second decade without having achieved the scale of systemic adoption that was once expected, it is worth asking whether these financial risks may in fact be larger than previously assumed, and whether they will be fairly distributed across sectors and groups.

A safer policy of investment might be one that accounts for the reality that AI development could continue to progress along a boom-and-bust cycle of growth. However, national AI strategies generally do not include measures to reduce the state's exposure to the financial losses that would be incurred if AI progress continues to fail in real-world use. This could be a particular concern for low-income states, which might over-expose themselves in risky AI while under-investing in other technologies or areas that may be essential for prosperity and sustainable growth, such as agricultural technologies, clean energy, sanitation and public health. (If anything, a state may only be able to truly enjoy the benefits of AI if it has first cleared other development milestones. For example it is unlikely that an AI system for better healthcare will be of universal benefit to a population where there is a lack of birth registrations or where non-male individuals are conferred fewer rights to government services.)

---

33 According to one study that is cited widely in AI policy, 'global GDP will be up to 14% higher in 2030 as a result of the accelerating development and take-up of AI – the equivalent of an additional $15.7 trillion'. See PwC (2017), 'Sizing the Prize', https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Using the PWC forecast and its own 'national statistics', the United Arab Emirates Strategy states that 'assuming automation happens to the full extent it can in each industry, there is a potential gain of AED 335 billion [approx. $91 billion] in increased economic output for the UAE'. United Arab Emirates Minister of State for Artificial Intelligence, Digital Economy & Remote Work Applications Office (2021), 'UAE National Strategy for Artificial Intelligence 2031', https://ai.gov.ae/strategy. PwC's projection also appears to be the basis for the Australian Department of the Prime Minister and Cabinet's (PMC) assertion, in a March 2022 paper, that 'it has been estimated that AI could contribute more than $20 trillion dollars to the global economy by 2030'. The PMC paper uses this claim as support for its argument that 'Australia needs to adopt these technologies in order to ensure that productivity and living standards keep pace with the rest of the world'. Department of the Prime Minister and Cabinet (2022), *Positioning Australia as a Leader in Digital Economy Regulations: Automated Decision Making and AI Regulation*, Canberra: Department of the Prime Minister and Cabinet, March 2022, pp. 1–2, https://consult.industry.gov.au/automated-decision-making-ai-regulation-issues-paper.

34 Yamashita, I., Murakami, A., Cairns, S. and Galindo-Rueda, F. (2021), 'Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative', OECD Science, Technology and Industry Working Papers 2021/09, pp. 9–10, https://www.oecd-ilibrary.org/science-and-technology/measuring-the-ai-content-of-government-funded-r-d-projects_7b43b038-en;jsessionid=nGJp7wXiL5BTYXmg5sTJb9dU.ip-10-240-5-158.

35 Ransbotham, S. et al. (2020), 'Expanding AI's Impact with Organizational Learning', MIT Sloan Management Review, 20 October 2020, https://sloanreview.mit.edu/projects/expanding-ais-impact-with-organizational-learning. A 2022 survey of business executives from a broad range of US industries found that 'implementation of AI into widespread production remains low'. See NVP (2022), *The Quest to Achieve Data-Driven Leadership: A Progress Report on the State of Corporate Data Initiatives*, NewVantage Partners, https://www.newvantage.com/_files/ugd/e5361a_2f859f3457f24cff9b2f8a2bf54f82b7.pdf.

36 Siegel, E. (2022), 'Models Are Rarely Deployed: An Industry-wide Failure in Machine Learning Leadership', KD nuggets, 17 January 2022, https://www.kdnuggets.com/2022/01/models-rarely-deployed-industrywide-failure-machine-learning-leadership.html.

Seeking a new term

The aspirational view of AI outlined in this section is so dominant in the discourse that even when policy language does not provide a concrete definition, the mere use of the term 'artificial intelligence' connotes this view of what the technology is (or is imagined to be). As a result, there have been calls in recent years to replace the term 'artificial intelligence' with specific terminology that foregrounds the technical realities of the systems in question and the parties responsible for its use.

For example, the Center on Privacy and Technology at Georgetown Law in Washington, DC, has announced that it will cease to use 'AI', 'artificial intelligence' and 'machine learning' in its work and will instead use specific terminology. Rather than stating that 'employers are using AI to analyze workers' emotions', the Center's staff will now use language such as: 'employers are using software advertised as having the ability to label workers' emotions based on images of them from photographs and video. We don't know how the labeling process works because the companies that sell these products claim that information as a trade secret.'[37]

Another potential approach is to develop follow-up questions to accompany mention of AI in the policy sphere. These could include questions like 'what type of AI?'; 'whose AI?'; 'who built this AI?'; 'was this AI built for this specific purpose?'; and 'is this AI deployed, under development, or at the concept stage?'.

37 Tucker, E. (2022), 'Artifice and Intelligence', Tech Policy Press, 17 March 2022, https://techpolicy.press/artifice-and-intelligence.

03
The lore of data

When it comes to AI, data is not always 'the new oil'. Often, datafication in the service of AI development has dubious benefits and concrete risks.

Assumption: A principal enabler of AI development and deployment is data. Therefore, states wishing to increase their AI capacity should endeavour to collect, consolidate and distribute the greatest volume of relevant data.

Counterpoint: Not all applications of AI will necessarily benefit from the collection, centralization and distribution of data. Furthermore, any data collection and distribution activity carries serious risks. In some cases, those risks may outweigh the anticipated gains the data might yield for AI development.

One of the most embedded AI assumptions today is that it is impossible to become successful in the pursuit of AI without possessing a lot of data.[38] National strategies often include specific measures to enhance the collection of data and to build centralized data infrastructures that serve to make datasets for AI development accessible to both public and private stakeholders.[39] In one much-cited AI index, state readiness for using AI in public services is graded, in part, on national 'data availability' – a compound metric based on factors such as the amount

38 Galindo, L., Perset, K. and Sheeka, F. (2021), *An overview of national AI strategies and policies*, Going Digital Toolkit Note, No. 14, OECD, pp. 10–11, https://goingdigital.oecd.org/data/notes/No14_ToolkitNote_AIStrategies.pdf.

39 United Kingdom Office for Artificial Intelligence (2022), 'National AI Strategy'; Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025', p. 49 and p. 71; President of the Russian Federation (2019), *On the Development of Artificial Intelligence in the Russian Federation*, p. 13; Villani (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*, p. 19; Polish Ministry of Digital Affairs (2020), 'Policy for AI Development in Poland from 2020', p. 32; Ministerio de Ciencia, Tecnología, Conocimiento e Innovación (2020), *Política Nacional de Inteligencia Artificial [National Artificial Intelligence Strategy]*, Santiago: Gobierno de Chile, pp. 32–33, https://www.minciencia.gob.cl/uploads/filer_public/bc/38/bc389daf-4514-4306-867c-760ae7686e2c/documento_politica_ia_digital_.pdf; Casados, D. et al. (2020), 'Agenda National Mexicana de Inteligencia Artificial' [Mexican National Agenda for Artificial Intelligence], p. 16, https://oecd.ai/en/wonk/documents/mexico-mexican-national-ai-agenda-2018-2030; see more in Van Roy et al. (2021), 'AI Watch - National strategies on Artificial Intelligence: A European perspective', p. 16.

of open data each government publishes and the level of mobile phone and internet use in the population (which is a proxy for how much digital data each citizen generates).[40]

However, this assumption belies a much more complex reality. The value of data for AI varies by application and depends upon the user's capacity for leveraging those data. Meanwhile, amassing and disseminating data can create risks and vulnerabilities that cannot necessarily be addressed through the privacy controls and security measures that states often promise as part of their AI campaigns.

The new oil?

It is often said that 'data is the new oil'. But the relationship between the availability of data and the performance of AI is far more fraught than the relationship between an entity's access to energy sources and its energy security. In reality, collecting data is much cheaper and easier than turning those data into an AI advantage.

In order to be useful for training most machine learning-based AI systems, data need to be well-curated and free from errors. Expunging errors from data is a non-trivial challenge.

In order to be useful for training most machine learning-based AI systems, data need to be well-curated and free from errors. Expunging errors from data is a non-trivial challenge.[41] Just as crucially, data must be closely aligned with the purpose for which the AI system is being developed. For example, if a machine learning system is to be used for medical triage or diagnostics, it must be trained on vetted historical patient data that have the same statistical properties as those of the patient population that it will be used on. It could not be trained on data from hospitals in another country. Even using data from a hospital in a different area of the same country might degrade the performance of the system.[42]

---

40 Oxford Insights (2021), 'Government AI Readiness Index 2021', https://www.oxfordinsights.com/government-ai-readiness-index2021; another report comparing AI capability in the US, China and the EU explains in its methodology that it uses data availability as a key indicator because 'more and higher-quality data will create new opportunities to use machine learning in AI applications'. See Castro, D. and McLaughlin, M. (2021), *Who Is Winning the AI Race: China, the EU, or the United States? – 2021 Update*, Center for Data Innovation, p. 4, https://www2.datainnovation.org/2021-china-eu-us-ai.pdf.
41 Kang, D. et al. (2022), 'Finding Errors in Perception Data With Learned Observation Assertions', Stanford Dawn, 24 January 2022, https://dawn.cs.stanford.edu/2022/01/24/loa.
42 Dexter, D. P., Grannis, S. J., Dixon, B. E. and Kasthurirathne, S. N. (2020), 'Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange', AMIA Joint Summits on Translational Science proceedings, AMIA Joint Summits on Translational Science, 2020, pp. 152–161, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233074.

Those data that are sufficiently representative at the time of collection never remain so indefinitely. Many contexts in which AI is deployed will evolve gradually (or sometimes not so gradually), a phenomenon known as 'distribution shift' that can significantly degrade the AI system's performance over time. Often, this happens in a way that is difficult to pre-empt.[43]

Nor can datasets ever be, as one strategy claims optimistically, a 'single platform of truth'.[44] Datasets only reflect the numerical approximation of a reality that may be far more multi-faceted. A standardized test score, for example, is an incomplete indicator of a student's total academic potential. Data also invariably harbour inconsistencies and biases.[45] As the tasks that the AI is intended for grow in complexity, the challenge of producing and maintaining clean, representative truthful data expands exponentially. A common refrain in the machine-learning community when discussing data is 'garbage in, garbage out'. But the evidence suggests that for a sufficiently complex AI task, all data are 'garbage'; they are naturally more limited, invariable, inflexible and biased than the reality that they purport to represent.[46]

Even when an embarrassment of what might appear to be well-matched data are available for an AI system, that does not guarantee success.[47] For example, though large language models are trained on unthinkably large volumes of data, they have shown themselves to be adept at generating false or divisive written media.[48] Nor does a representative dataset always necessarily lead to AI that generates positive outcomes; a dataset of language from unmoderated internet forums, for example, may be representative of what people say in those spaces, but a chatbot trained on such speech would be undesirable.[49]

---

43 Shendre, S. (2020), 'Model Drift in Machine Learning', Towards Data Science, 14 May 2020, https://towardsdatascience.com/model-drift-in-machine-learning-models-8f7e7413b563; Sculley, D. et al. (2014), 'Machine Learning: The High Interest Credit Card of Technical Debt', *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, https://research.google/pubs/pub43146.
44 Saudi Data and AI Authority (2020), *Realizing Our Best Tomorrow: Strategy Narrative*, p. 12, https://www.carringtonmalin.com/wp-content/uploads/2020/08/NDAIS-Strategy-Narrative-V2-19Oct20.pdf.
45 Simply eliminating bias from these datasets, as many strategies propose, is not a viable proposition if the structural inequalities that underpin them are deeply embedded; see Knight, H. E. et al. (2021), 'Challenging racism in the use of health data', *The Lancet Digital Health*, 3, 4, E144-146, 1 March 2021, https://doi.org/10.1016/S2589-7500(21)00019-4.
46 'Garbage in, garbage out' – another popular AI aphorism for describing data's relationship to model quality – is also not particularly helpful; it implies that the opposite – non-garbage in, non-garbage out – is also true. But models built on good data can also fail.
47 Sherman, J. and Sacks, S. (2019), 'The Myth of China's Big A.I. Advantage', Slate Future Tense, 13 June 2019, https://slate.com/technology/2019/06/data-not-new-oil-kai-fu-lee-china-artificial-intelligence.html.
48 Buchanan, B., Lohn, A., Musser, M. and Sedova, S. (2021), 'Truth, Lies, and Automation: How Language Models Could Change Disinformation', Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/truth-lies-and-automation.
49 Fingas, J. (2022), 'AI trained on 4chan's most hateful board is just as toxic as you'd expect', engadget, 8 June 2022, https://www.engadget.com/ai-bot-4chan-hate-machine-162550734.html.

The perils of datafication

Because an imagined application of AI[50] can only be tested if data relating to that application are available, digitized and consolidated,[51] the pursuit of AI is seen to require the mass 'datafication' of society. Or, as the scholars Ulises A. Mejias and Nick Couldry put it, 'the transformation of human life into data through processes of quantification, and the generation of different kinds of value from data'.[52] Though the potential benefits of such datafication are widely discussed in national AI strategies, its inherent risks receive less attention.

Every time information relating to people is turned into machine-readable data, it creates new privacy risks. Indeed, many of the characteristics that will make a dataset suitable for building AI will be particularly bad for the privacy of the people represented within those data. Machine learning-based AI often thrives when trained on massive, granular, multi-modal, labelled data that can reveal sensitive personal information even when individual datapoints are anonymized.

The risks of such AI are multiplied whenever data are consolidated in the types of national strategic data repositories that many states are seeking to build as a foundation for their AI campaigns.[53] Especially in countries lacking rigorous data protection regimes, such repositories[54] could afford authorities and nefarious actors easy access to personal information that would have previously been siloed in separate streams. This creates novel possibilities for abuse.

Of course, these dataset and data infrastructures could be built and managed according to strict privacy principles, as many strategies note. However, privacy controls can always be revoked. A malevolent new regime that takes power could abuse data that were previously collected and used in tight compliance with privacy protections. A shifting landscape of criminal law can also have a direct effect on the risk that these datasets pose to the people they contain. Data that reveal individuals to have engaged in activities that were once legal (such as seeking an abortion) may become problematic if those activities are suddenly made illegal.

Abuse by private actors is also a concern. Datasets that are made available to a diversity of stakeholders can become 'runaway datasets' that are so widely held, stored, distributed and reproduced that they cannot be recalled if they are discovered to be problematic.[55] Such runaway datasets are already common, and their risks expand the longer they are accessible and as their scale and diversity

---

50 For example, as one strategy proposed, using AI chatbots to 'comfort earthquake survivors': see Kumar et al. (2018), *National Strategy for Artificial Intelligence #AIforAll*, p. 15.
51 In the case of the chatbot, this might include digitized transcripts of conversations between earthquake survivors and emergency personnel. Such transcripts would likely include sensitive information.
52 Mejias, U. A. and Couldry, N. (2019), 'Datafication', *Internet Policy Review*, 8(4), doi.10.14763/2019.4.1428.
53 OECD (2019), 'Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies', https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en.
54 As one strategy put it, the kind of repository that would result from 'an aggressive policy aimed at promoting data access'. Villani (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*, p. 19.
55 Peng, K., Mathur, A. and Narayanan, A. (2021), *Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers,* https://arxiv.org/pdf/2108.02922.pdf.

increase.[56] The longer a dataset is available, the greater the risk that it will also end up being used for problematic or unproven applications. This has already been observed in the research sector. In one case, the US technology firm Microsoft took down a publicly available dataset of millions of images of more than 100,000 'celebrities' (many of whom were journalists) after it was revealed that it had been used by a number of companies that build surveillance technologies, including IBM, Panasonic, Hitachi, Sensetime and Megvii.[57]

Finally, datafication can increase the risk of cybercrime. Even closed data repositories that are only made available on a restricted basis can be hacked. Such attacks could make potentially sensitive data available for crimes such as identity theft or stalking. Large breaches could also potentially enable malign actors to 'poison' these datasets, so that any AI systems trained upon them could exhibit suboptimal or dangerous performance.[58] The more that a dataset is disseminated, the higher the chance that it could be attacked.

The datafication mindset

An emphasis on data availability for AI might stand in the way of rigorous privacy protections for preventing the kinds of harms described above. Many observers have noted that authoritarian regimes will have greater access to data for AI development than societies where the mass collection of personal information is subject to controls.[59] While this has been helpful for illuminating the perils of mass datafication, there is a risk that it could serve arguments that undermine the push for better digital privacy.[60] In a 2021 report comparing the AI capabilities of the US, the EU and China, one think-tank went so far as to argue that, '[US] Congress should ensure any change to federal data privacy legislation does not limit data collection and use of AI'.[61]

Datafication in the service of AI could also have profound secondary effects. Just as misinformed visions of the actual 'intelligence' of AI could cause governments to prioritize it over other areas of government investment (see previous chapter), an over-emphasis on data could privilege research efforts focused on large models

---

56 Researchers at Facebook found in 2021 that the company has no way of keeping track of all of the personal user data that it has collected over the years since its founding. See Franceschi-Bicchierai, L. (2022), 'Facebook Doesn't Know What It Does With Your Data, Or Where It Goes: Leaked Document', Motherboard, 26 April 2022, https://www.vice.com/en/article/akvmke/facebook-doesnt-know-what-it-does-with-your-data-or-where-it-goes.
57 Harvey, A. and LaPlace, J. (2021), 'MS-CELEB-1M', exposing.ai, https://exposing.ai/msceleb; Madhumita, M. (2019), 'Microsoft quietly deletes largest public face recognition data set', *Financial Times*, 6 June 2019, https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2.
58 Pupillo, L., Fantin, S., Ferreira, A. and Polito, C. (2021), *Artificial Intelligence and Cybersecurity: Technology, Governance and Policy Challenges*, Brussels: Centre for European Policy Studies, https://www.ceps.eu/ceps-publications/artificial-intelligence-and-cybersecurity-2.
59 As Schmidt and Allison argue, elements that bolster China's 'surveillance state', including 'government, laws and regulations, public attitudes about privacy, and thick cooperation between companies and their government are all green lights for its advance of AI'. See Allison, G. and Schmidt, E. (2020), 'Is China Beating the U.S. to AI Supremacy?', Belfer Center for Science and International Affairs, https://www.belfercenter.org/publication/china-beating-us-ai-supremacy; see also Castro and McLaughlin (2021), *Who Is Winning the AI Race: China, the EU, or the United States? – 2021 Update*; Filgueiras, F. (2022), 'The politics of AI: democracy and authoritarianism in developing countries', *Journal of Information Technology & Politics*, doi.10.1080/19331681.2021.2016543; Stavridis, J. (2021), 'Artificial Intelligence is America's Achilles Heel against China', Bloomberg.com, 20 May 2021, https://www.bloomberg.com/opinion/articles/2021-05-20/china-s-artificial-intelligence-advantage-is-america-s-achilles-heel.
60 Sherman and Sacks (2019), 'The Myth of China's Big A.I. Advantage'.
61 Castro and McLaughlin (2021), *Who Is Winning the AI Race: China, the EU, or the United States? – 2021 Update.*

(i.e. highly complex machine learning-based AI systems with many parameters). Unlike simpler tools, these large models have increasingly proven to be problematic in terms of both their societal and environmental harms.[62]

By extension, the massive provision of data to industry could encourage a misplaced focus on data-driven approaches to problems that are rooted in societal issues. It is tempting to assume that any societal challenge can be solved with scientific exactitude by training a machine-learning model on that challenge. But datafying a particular problem might only succeed in giving rise to an inflexible machine-learning model that provides no true insight or predictive capacity in operational use – at worst, such efforts become a costly distraction from solutions that could have a much higher long-term probability of success.

Even if an AI were to exhibit some positive performance in tackling a difficult challenge, it might not be enough to justify the risks associated with the necessary datafication. Consider, for example, AI efforts for suicide prevention, or for detecting welfare fraud; even if they were to create some efficiencies in tackling these noble causes, such initiatives require the collection and dissemination of highly sensitive data that could be breached or abused to harmful effect. This is not to say that reducing suicides or fraud are not goals that states should pursue. Rather, if similar gains can be achieved by measures that do not involve the collection of data, those measures would be preferable to ones that do.

Nor can one necessarily expect the benefits of datafication to be distributed evenly among stakeholders.[63] Mass datafication naturally privileges communities and sectors with the capacity and computational resources to process data – such as the tech sector and high-revenue industries such as finance. Meanwhile, mass datafication will yield few gains for those who do not have access to the necessary resources to protect themselves from the collection of information that can be used against their interests.[64]

State AI strategies have yet to grapple fully with these concerns.[65] It might therefore be helpful for states to assume that any datafication in the service of AI carries tangible risks that not only must be weighed against the potential benefits of the resulting AI capabilities but also against the very real possibility that those capabilities will fail. In this analysis, some states may find that some of their proposed datafication measures simply do not justify the risk. As always, the buy-in of those groups who will be most affected by this datafication is a key criterion in this analysis.

---

62 Bender, E. M., Gebru, T., McMilla-Major, A. and Shmitchell, S. (2021), 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency: 610–623, p. 612, https://doi.org/10.1145/3442188.3445922; Patterson, D. et al. (2021), 'Carbon Emissions and Large Neural Network Training', https://arxiv.org/pdf/2104.10350.pdf.
63 GPAI (2022), 'Data Governance Working Group – A Framework Paper for GPAI's Work on DataGovernance 2.0', Report, November 2022, Paris: Global Partnership on AI https://gpai.ai/projects/data-governance/gpai-data-governance-wg-report-2022.pdf.
64 Goben, A. and Sandusky, R. J. (2020), 'Open data repositories: Current risks and opportunities', *College & Research Libraries News* 81(2): p. 62, https://crln.acrl.org/index.php/crlnews/article/view/24273/32092and.
65 For example, according to UNICEF, even among AI strategies that do address privacy there is a troubling lack of specific measures to protect the rights of the single demographic arguably most vulnerable to intrusion and abuse: children. See Penago, M. (2020), 'What do national AI strategies say about children? Reviewing the policy landscape and identifying windows of opportunity', UNICEF, https://www.unicef.org/globalinsight/stories/what-do-national-ai-strategies-say-about-children.

04
A race with no winners

The AI 'race' is a misguided metaphor, if acted upon too literally, it could stand in the way of robust, fair policy.

Assumption: In order to succeed in international power competition, states must develop and deploy AI more widely and more quickly than their adversaries and peers.

Counterpoint: A race-like approach to technology development could stand at odds with a state's capacity to adopt AI in a way that truly serves the common good.

It is a received truth that AI will be a fundamental pillar of economic, military and geopolitical power in the 21st century. The stakes of achieving such 'AI supremacy' are not only seen to be inconceivably rich,[66] but also existential. And because AI is generally characterized as a fast-moving technology, the AI power discourse tends to subscribe to the view that the technology's richest spoils are reserved for those who move first and fastest.[67] Put another way, AI is seen as a race with winners and losers.[68]

---

66 PwC (2017), 'Sizing the Prize'.

67 National strategies often couch their proposed efforts in terms of the need to claim the technology's 'first mover's advantage'. A review of eight Nordic AI strategy documents found that almost all referred to a 'first-mover advantage'. See Dexe, J. and Franke, U. (2020), 'Nordic lights? National AI policies for doing well by doing good', *Journal of Cyber Policy* 5(3): pp. 332–349, https://doi.org/10.1080/23738871.2020.1856160; State Council (2017), 'A New Generation Artificial Intelligence Development Plan', translation by Webster, G., Creemers, R., Triolo, P. and Kania, E., https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017, quoted in Cave, S. and ÓhÉigeartaigh, S. S. (2018), 'An AI Race for Strategic Advantage: Rhetoric and Risks', *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*: pp. 36–40, https://doi.org/10.1145/3278721.3278780; Saudi Data and AI Authority (2020), *Realizing Our Best Tomorrow: Strategy Narrative*, p. 25.

68 It is not uncommon for analysis and commentary to go so far as to declare 'winners' and 'losers' in the AI race, particularly in comparisons of near-peer superpowers such as the US, China, Russia and the EU. See Reuters (2021), 'China has won AI battle with U.S., Pentagon's ex-software chief says', 11 October 2021, https://www.reuters.com/technology/united-states-has-lost-ai-battle-china-pentagons-ex-software-chief-says-2021-10-11; Cooper, J. and Kompella, K. (2022), 'No, China is not winning the AI race', The Hill, 3 February 2022, https://thehill.com/opinion/technology/592270-no-china-is-not-winning-the-ai-race; Berggruen, N. and Gardels, N. (2018), 'A wakeup call for Europe', *Washington Post*, 27 September 2018, https://www.washingtonpost.com/news/theworldpost/wp/2018/09/27/europe; Allen, J. and Husain, A. (2017), 'The Next Space Race Is Artificial Intelligence', *Foreign Policy*, 3 November 2017, https://foreignpolicy.com/2017/11/03/the-next-space-race-is-artificial-intelligence-and-america-is-losing-to-china.

In the private sector, it is true that those companies that can roll out a technology quicker than competitors will enjoy a first-mover advantage. And yet when it comes to state efforts, this dynamic is a little more complex. While data, investment, a large AI workforce and permissive regulations may enable a state to have more AI than its peers, given AI's persistent limitations in critical functions, there is no guarantee that this would result in a net strategic advantage. Nor would more AI adoption necessarily result in more equitable benefits for a state's population compared to those enjoyed by the citizens of a state with less AI. Meanwhile, the competitive pursuit of AI supremacy could result in a potentially profound misalignment between AI policy and the core principles of responsible technology governance. The 'AI race' is therefore an imperfect metaphor, and one that potentially poses risks if acted upon too literally.

The risks of an AI race

In modern technology discourse, it is taken as read that the pace of technological change has surpassed the pace of rule-making. It is further assumed that innovation can be stifled by overzealous or excessive regulations.[69] But is this a prudent approach?

In the context of the race for AI supremacy, a race-like mindset inevitably serves those calling for states to accelerate, or even fundamentally rethink, the rule-making process.

In the context of the race for AI supremacy, a race-like mindset inevitably serves those calling for states to accelerate,[70] or even fundamentally rethink, the rule-making process.[71] This is problematic. While it is true that AI challenges the traditional rule-making process, there are no clear blueprints for the length of time that it should take to comprehensively regulate such a complex technology with such widespread applications. Even if there were a way to create overnight AI laws, it is unclear how much of today's AI is sufficiently predictable and traceable to be effectively governed by rules on, say, fairness, accountability and safety (see Chapter 5). It is therefore possible that AI rule-making is being held to an unrealistic standard.

---

69 Russia's strategy states that 'excessive regulation in this sphere might significantly slow the pace of development and introduction of technological solutions', see President of the Russian Federation (2019), *On the Development of Artificial Intelligence in the Russian Federation.* The EU AI draft bill states that 'rules for artificial intelligence should be balanced, proportionate and not unnecessarily constrain or hinder technological development', see 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'. Also, see Polish Ministry of Digital Affairs (2020), 'Policy for AI Development in Poland from 2020', p. 23.

70 Steering group and secretariat of the Artificial Intelligence Programme (2019), *Leading the way into the era of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019*, p. 44; Casados (2020), 'Agenda National Mexicana de Inteligencia Artificial', p. 104.

71 Steering group and secretariat of the Artificial Intelligence Programme (2019), *Leading the way into the era of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019*, p. 44; Casados (2020), 'Agenda National Mexicana de Inteligencia Artificial', p. 104.

More fundamentally, this view risks framing regulation and innovation as opposing interests.[72] In the context of an AI race, the possibility that a competitor might get ahead is a powerful argument for, say, exempting a given venture from regulations.[73] States competing to make their regulatory environments more welcoming for AI[74] could take regulatory risks at the expense of safety and fairness.[75] This could put a state's AI strategy at odds with the interests of its population – especially the interests of those who are most likely to experience harms. At worst, the AI race pits a state's (possibly fictive) mandate to stay ahead of its competitors in the development and use of AI against its real responsibility to protect its most vulnerable citizens from harmful technology.

A focus on moving faster than one's adversaries has also marred multilateral efforts in AI governance. The Convention on Certain Conventional Weapons, for example, has failed to develop binding rules for autonomous weapons in part because of a number of states' concerns that such rules might stand in the way of the development of national technological capacity.

Box 1. 'Sandboxes' and other non-traditional policy instruments

A number of states and other stakeholders have proposed 'sandboxes' – an instrument that allows companies to pilot limited operations for a set period of time outside of the existing regulatory framework – as a potentially safe compromise between the imperatives of rigorous regulation and competitiveness.[76] Sandboxes are potentially useful, but they are not a panacea. Given AI's sensitivity to context (an AI that works well in one scenario might fail in a slightly different one), the results of a sandboxing experiment do not necessarily indicate the likely outcomes if that experiment were to be replicated at scale.[77] Furthermore, while sandbox schemes have been trialled fairly extensively in other sectors, their use in AI remains mostly 'embryonic', according to one report,[78] and have not yet been extensively examined

---

72 Truby, J., Brown, R. D., Ibrahim, I. A. and Parellada, O. C. (2021), 'A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications', *European Journal of Risk Regulation*, 13(2), pp. 1–29, doi:10.1017/err.2021.52.
73 For example, the EU AI draft bill proposes that 'under exceptional reasons… Member States could authorise the placing on the market or putting into service of AI systems which have not undergone a conformity assessment'. See 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', p. 66. France's AI strategy similarly proposes 'temporary lifting of certain regulatory constraints in order to leave the field free for innovation'. See Villani (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*, p. 47.
74 For example, Saudi Arabia's AI strategy includes the goal of 'enact[ing] the most welcoming legislation', for AI development and deployment. Saudi Data and AI Authority (2020), *Realizing Our Best Tomorrow: Strategy Narrative*, p. 22.
75 Cave and ÓhÉigeartaigh (2018), 'An AI Race for Strategic Advantage: Rhetoric and Risks'.
76 President of the Russian Federation (2019), *On the Development of Artificial Intelligence in the Russian Federation*; Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025'; 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'; Polish Ministry of Digital Affairs (2020), 'Policy for AI Development in Poland from 2020'; Casados (2020), 'Agenda National Mexicana de Inteligencia Artificial'.
77 Pop, F. and Adomavicius, L. (2021), 'Sandboxes for Responsible Artificial Intelligence', European Institute of Public Administration, https://www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence. For an example of a real sandbox experiment, see Secure Project's trial of a workplace profiling AI system to rank employees by their cybersecurity risk. Datatisynet (2021), 'Secure Practice – sluttrapport', https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/secure-practice---sluttrapport.
78 Van Roy et al. (2021), 'AI Watch – National strategies on Artificial Intelligence: A European perspective', p. 15.

in the literature.[79] Nor can industry be expected to take the lead on self-regulation; as a team of researchers from Google, OpenAI and Berkeley wrote in early 2022, 'the competitive dynamics' of the contemporary AI field may pressure companies 'to take shortcuts on safety' regardless of whether they have direct economic incentives to make safe systems.[80]

Describing AI integration as a race may be an acceptable turn-of-phrase at the strategy stage of policymaking, but it could lead to trouble when it comes to implementing the kinds of structural measures that AI ethics likely require. An aggressive race-like posture could even lead states to side-step certain ethical precepts entirely, if securing an advantage in a particular domain is deemed to be sufficiently urgent.[81] (Or, at a minimum, it risks reframing the application of AI ethical principles as a necessary stepping-stone to get ahead in the AI race[82] – rather than an end in itself.)

Running, not talking

The race for AI emphasizes a narrow, predominantly Western vision of technological progress that valorizes raw scale,[83] and disregards other indicators that might be more accurately indicative of a state's capacity to adopt AI in a way that truly serves the common good. In an alternative framing of the AI race, critical factors may include openness and transparency of institutions and freedom of civil society and the press (which would be key for the implementation of 'accountable' AI), rule of law and economic equality (which are necessary for AI projects to successfully engender fair outcomes), and educational attainment in other fields outside of science, technology, engineering and mathematics (which would bolster a state's capacity to untangle the complex philosophical and legal challenges posed by the automation of critical functions).

And yet, it is difficult to have this type of conversation because the notion itself of an AI race stands at odds with the kind of truly inclusive discursive process that is likely to be warranted for the responsible implementation of technology

---

79 Truby, Brown, Ibrahim and Parellada, (2021), 'A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications'.
80 Hendrycks, D., Carlini, N., Schulman, J. and Steinhardt, J. (2022), 'Unsolved Problems in ML Safety', https://arxiv.org/pdf/2109.13916.pdf.
81 For example, employing a novel autonomous military system would likely require structural changes to long-standing test and evaluation protocols – a process that could take precious years to implement. See Haugh, B. et al. (2018), 'The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems', Washington: Institute for Defense Analyses, September 2018, https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx. In a similar vein, Finland's AI strategy questions whether it is necessary for deep neural network-based systems to be explainable: see Steering group and secretariat of the Artificial Intelligence Programme (2019), *Leading the way into the era of artificial intelligence Final report of Finland's Artificial Intelligence Programme 2019*, p. 36.
82 Daly, A. et al. (2021), 'AI, Governance and Ethics: Global Perspectives', in Micklitz, H. et al. (eds) (2021), *Constitutional Challenges in the Algorithmic Society* (pp. 182–201), Cambridge: Cambridge University Press, doi:10.1017/9781108914857.010; Dexe and Franke (2020), 'Nordic lights? National AI policies for doing well by doing good'.
83 For example, the number of new AI startups, the number of AI patents, the number of graduates in AI-relevant fields, the level of public spending in AI, the number of ethics boards, the number of bills, the number of job fairs, of regulatory sandboxes, of pilot programmes.

that serves everyone equally. The assumption that AI development is a race leaves little breathing room for any arguments that question the race itself – or arguments that question, for example, whether winning the race in this particular way would be a universal good.

A closer look at AI indexes

When states want to see their progress in the AI race, they often turn to an AI index. Over the past several years, a number of indexes, which rank states against one another in terms of their relative 'readiness' or 'capacity' for AI development and adoption,[84] have become a fixture in the AI governance sphere. For many states, rising in the rankings has even become a matter of policy[85] (some of the indexes themselves actively encourage a race-like approach to AI policy[86]). Yet there is evidence to suggest that these indexes are an imperfect scorecard of relative national progress in AI.

First, AI 'readiness' or 'capacity' are ambiguous concepts that can only be gauged by inconsistent, divergent proxy indicators. The Oxford Insights index ranks countries based on just 10 indicators that are given equal weight despite representing vastly different national characteristics. For example, whether a country has an AI strategy is weighted equally to the national rate of internet usage. Seven of these indicators are derived directly from other indexes,[87] which further confounds evaluation.

---

84 Zhang, D. et al. (2022), 'The AI Index 2022 Annual Report', AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, https://aiindex.stanford.edu/report; Fuentes Nettel, P. et al. (2021), 'Government AI Readiness Index 2021', Oxford Insights, https://www.oxfordinsights.com/government-ai-readiness-index2021; Tortoise Media (2021), 'The Global AI Index', https://www.tortoisemedia.com/intelligence/global-ai.

85 It is not uncommon for state strategies to cite indexes as the primary indicator of their progress in certain dimensions of AI adoption. See, for example, Presidency of the Republic Türkiye (2021), 'National Artificial Intelligence Strategy 2021-2025', pp. 20–25; United Arab Emirates Minister of State for Artificial Intelligence, Digital Economy & Remote Work Applications Office (2021), 'UAE National Strategy for Artificial Intelligence 2031'; Saudi Data and AI Authority (2020), *Realizing Our Best Tomorrow: Strategy Narrative*, p. 28. Numerous states have even targeted a specific ranking as the goal of their AI strategies – for example, Israel aims to become 'one of the top five countries in the world in AI', see Hennessey, Z. (2022), 'Israel's critical role in the future of AI', *Jerusalem Post*, 8 February 2022, https://www.jpost.com/business-and-innovation/article-695865; Poland aims to rank 'among the top ten countries in the AI Readiness Index', see Polish Ministry of Digital Affairs (2020), 'Policy for AI Development in Poland from 2020', p. 28; Australia seeks to be 'a top 10 digital economy and society by 2030', see Department of the Prime Minister and Cabinet (2022), *Positioning Australia as a Leader in Digital Economy Regulations: Automated Decision Making and AI Regulation*; Argentina hopes to be the 'most advanced country in AI' in the region, see Presidencia de la Nación Argentina (2020), *Plan Nacional de Inteligencia Artificial*, Buenos Aires, p. 224, https://ia-latam.com/wp-content/uploads/2020/09/Plan-Nacional-de-Inteligencia-Artificial.pdf; Ministerio de Ciencia, Tecnología, Conocimiento e Innovación (2020), 'Política Nacional de Inteligencia Artificial', p. 5, https://www.minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial.

86 The Stanford Global AI Vibrancy Tool, which is part of one of the most comprehensive and widely cited AI indexes, characterizes itself as an instrument to see 'who is leading the global AI race'. Stanford HAI (2022), 'Global AI Vibrancy Tool', Stanford Institute for Human-centered AI, Stanford University, https://aiindex.stanford.edu/vibrancy. In 2019 the authors of the Oxford Insights Government Artificial Intelligence Readiness Index did acknowledge in a methodology annex that 'there is a risk that indices such as these create a global race for AI', see Miller, H. et al. (2019), 'Government Artificial Intelligence Readiness Index 2019', Oxford Insights, p. 6, https://www.oxfordinsights.com/ai-readiness2019. (The authors appear to have removed this note in the 2020 and 2021 editions.)

87 Fuentes Nettel et al. (2021), 'Government AI Readiness Index 2021', p. 70.

Meanwhile, some of the individual indicators commonly used in indexes are potentially inappropriate for any kind of like-for-like comparison between states. For example, a metric of each country's public investments in AI[88] cannot account for the potentially very significant differences in how those investments are spent.[89] Indexes that rank states by their number of 'AI players' do not distinguish between very large and prolific organizations such as a leading research university and small ventures that are much less likely to produce novel AI.[90] Similarly, metrics relating to the number of 'AI projects' or 'AI services' (for example, public administration functions or private sector products that involve AI) do not always differentiate between types of AI. Indexes may, for example, count a large search service that uses cutting-edge natural language processing the same way they count a city council that uses software with a decades-old algorithm.

Some of the key metrics used in these rankings also appear to draw from unreliable data. The Stanford Human-centred Artificial Intelligence (HAI) Index's ranking of 'Relative AI Skill Penetration Rate by Geographical Area' and more than a dozen indicators used in Tortoise Media's Global AI Index rely on a LinkedIn dataset derived largely from information that is self-added by LinkedIn users,[91] some of whom inevitably make distorted claims about their AI-relevant skills. Indicators of AI adoption by companies also rely on self-reporting that might be prone to inaccuracies. Stanford's ranking of 'AI Adoption by Organizations in the World', is derived from a voluntary online survey of several thousand executives conducted by McKinsey & Company – a survey whose detailed methodology and raw data are not publicly released.[92]

There is also evidence of potential regional and demographic biases in both index data and development. For example, raw data on journal citations, a key indicator in several indexes, can under-represent the academic output of institutions and practitioners from the Global South. (There is evidence to suggest that biases can be a major factor contributing to the under-representation of Global South authors, articles and journals in citation counts.)[93] LinkedIn, a key source of data, does not have the same user rates all over the world, meaning that it may under-represent AI penetration in certain regions. Meanwhile, a disproportionate number of these data and indicators are compiled, sorted and presented by Western institutions,[94] and non-male experts are under-represented in some indexes' staffing and advisory bodies.[95]

---

88 For example, Tortoise Media (2021), 'The Global AI Index'.
89 For example, more than 80 per cent of US public investment in AI for 2022 was earmarked for defence and security, a vastly higher proportion than that of, say, Estonia's or Argentina's public investment in AI. See Zhang (2022), 'The AI Index 2022 Annual Report', p. 189.
90 Righi, R. et al. (2021), 'AI Watch Index 2021', Seville: Joint Research Centre, European Commission, https://ai-watch.ec.europa.eu/publications/ai-watch-index-2021_en.
91 Zhang et al. (2022), 'The AI Index 2022 Annual Report', p. 149; Tortoise Media (2021), 'The Global AI Index'.
92 Balakrishnan, T., Chui, M., Hall, B. and Henke, N. (2020), *The state of AI in 2020*, McKinsey & Company, https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020.
93 See Skopec, M., Issa, H., Reed, J. and Harris, M. (2020), 'The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis', *Research Integrity and Peer Review* (5)2, https://doi.org/10.1186/s41073-019-0088-0.
94 For example, all of the 'Research and Development' indicators in the Stanford Global AI Vibrancy Tool are based on data from the Center for Security and Emerging Technology at Georgetown University; three of its main 'Global AI Vibrancy Tool' indicators are compiled by LinkedIn, a US company.
95 For example, of the Tortoise Media Index's 26 advisers, only five appear not to be men, and of the 11 members of the Stanford AI Index 2021 report's steering committee, only two appear not to be men.

05
Mechanical ethics

Achieving ethical AI by narrow, technical means alone is an illusory goal. Broader societal measures are necessary to ensure that AI does not cause harm.

Assumption: Ethical principles can be encoded into AI.

Counterpoint: Achieving 'ethical AI' requires expansive measures that extend far beyond strictly technical fixes, including – potentially – uncomfortable organizational and societal reform.

States rightly recognize that it is impossible to implement AI successfully if they cannot set and enforce ethical principles.[96] However, this assumes that ethical AI, as commonly envisioned, is feasible. In fact, the 'translation to actions' of ethical AI principles, as the authors of one report put it, 'is often not obvious'.[97] Certain ethical AI principles likely require measures to address not only issues with the technology itself but also with those who develop, govern and use it. This means that organizations and states that do not act equitably, transparently or accountably are unlikely to be equipped to embed those principles in the design, development, regulation and use of AI. True ethical AI might therefore require lengthy, systemic reform. All of which raises uncomfortable questions about whether ethical principles can be reconciled with the precepts of AI supremacy. A failure to engage with these questions risks derailing the enterprise of AI ethics entirely.

96 For our purposes, 'ethical principles' here also include legal principles.
97 Brundage, M. et al. (2020), 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', https://arxiv.org/abs/2004.07213.

The technical challenges of AI ethics

The common principles of AI ethics[98] are inarguably noble. Few would claim that it would be bad to have a high level of 'fairness', 'safety', 'accountability' and 'transparency' in AI. However, as is clear from the many instances of complex AI failure in recent years,[99] as well as what we know to be inherently true of complex algorithmic systems, it is still far from certain that the technology *itself* can ever meet most states' definition of ethical AI.

For example, it is common for policies to call for critical AI to be 'explainable,' meaning that they can be understood by those who interact with them (be they users or subjects of the system). And yet, the creation of explainable high-performance AI, particularly deep-learning models, remains an open research challenge – perhaps even a mathematical impossibility.[100] Much AI that is marketed today as being 'explainable' does not actually meet this criterion in deployment; if it does, that explainability may very well be ineffective at enabling users to understand the system to an extent required by law, even in cases where those users are experts. Explainability may also often come at the expense of system performance.[101] Meanwhile, it would be difficult to set broad standards for system understandability, given that every user's capacity for understanding these systems will be unique. (None of which is a reason to abandon explainability – but rather a good reason to caveat any mention of explainability as a fix-all for AI understandability issues.)

The creation of explainable high-performance AI, particularly deep-learning models, remains an open research challenge – perhaps even a mathematical impossibility

Similarly, strategies that call for 'fair' or 'unbiased' AI assume that datasets and models can be expunged of bias. However, in any AI system where the data and the model do not perfectly represent the characteristics and dynamics of the environment in which they are used, bias can only be reduced.[102] The non-expungeable biases that persist cannot always be easily quantified. This is especially true when the population with which an AI system interacts is constantly shifting. A dataset that originally exhibited no evident harmful biases may become much less representative if its target population experiences demographic or economic

---

98 Jobin, A., Ienca, M. and Wayena, E. (2019), 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence* 1: pp. 389–99, https://www.nature.com/articles/s42256-019-0088-2; Fjeld, J. et al. (2020), 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI', Berkman Klein Center Research Publication No. 2020-1, http://dx.doi.org/10.2139/ssrn.3518482.
99 AI Incident Database (undated), 'Incident List', https://incidentdatabase.ai/summaries/incidents?lang=en.
100 Yampolskiy, R. V. (2019), 'Unexplainability and Incomprehensibility of Artificial Intelligence', https://arxiv.org/abs/1907.03869; d'Aquin, M. (2021), 'On the Impossibility of Explaining AI', Towards Data Science, 30 November 2021, https://towardsdatascience.com/on-the-impossibility-of-explaining-ai-aa0b39768375.
101 Ghassemi, M., Oakden-Rayner, L. and Beam, A. L. (2021), 'The false hope of current approaches to explainable artificial intelligence in health care', *Lancet Digital Health*, 3: e745-50, https://www.thelancet.com/action/showPdf?pii=S2589-7500%2821%2900208-9.
102 Balayn, A. and Gürses, S. (2021), 'Beyond Debiasing: Regulating AI and its inequalities', European Digital Rights, pp. 31–32, https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution.

changes. This poses a challenge for drawing a measurable and enforceable regulatory line between acceptable and unacceptable levels of bias, not to mention for developing measures to offset these biases.

Meanwhile, ethical principles such as reliability and predictability cannot be technologically assured. No matter how large they are, datasets and models can only capture historical trends, patterns, phenomena and statistical distributions.[103] Therefore it is rarely possible to guarantee that an AI system operating in an open, complex environment will not encounter inputs for which it is ill equipped to respond reliably or predictably.[104] Even an approximated litmus test for the performance of a machine-learning tool would depend on validation datasets that are, as the EU's proposed AI bill put it, 'relevant, representative, free of errors and complete'.[105] Such datasets for testing and validation could be beset by the same challenges that face the datasets on which systems are developed.

Nor is it clear that any amount of finite testing could identify all the ways that a system will experience either unforced errors or failures that are the result of intentional misuse by users.[106] Here, too, trying to gauge a system's unreliability or unpredictability (to determine whether it is above or below an acceptable threshold) is a challenge, since it is hard to develop concrete metrics for the degree of uncertainty regarding that system's future behaviour. Rigorous AI regulations would likely hinge on a novel scheme of recursive review that goes beyond the existing processes that are currently in place for, say, new weapon systems.[107]

Meanwhile, although third-party auditing is gaining traction as a potentially vital instrument for verifying whether an AI system is ethically aligned, the architecture for such audits remains skeletal.[108] Gaps also remain in the processes for ensuring that audits result in organizations reforming their algorithms or

---

103 Birhane, A. (2021), 'The Impossibility of Automating Ambiguity', *Artificial Life* 27(1), pp. 44–61, https://doi.org/10.1162/artl_a_00336. Even very high-performing machine-learning systems have shown a susceptibility to the effects of gradual shifts in the properties of the people or things that they model (for example, a system designed to optimize energy grids on the basis of energy consumption will see its performance degrade if patterns of energy use change over time). Such differences and shifts are difficult to detect in advance of their causing a system's performance to degrade. See Shendre (2020), 'Model Drift in Machine Learning'; Sculley (2014), 'Machine Learning: The High Interest Credit Card of Technical Debt'.
104 All systems granted a degree of autonomy are at risk of encountering 'edge cases' that fall outside the total scope of data that they were trained and tested on, and they have a likelihood of failing in such cases (this is a common characteristic of present-day AI systems known as 'brittleness'). See Cummings, M. L. (2017), *Artificial Intelligence and the Future of Warfare*, Research Paper, London: Royal Institute of International Affairs, https://www.chathamhouse.org/2017/01/artificial-intelligence-and-future-warfare; Holland Michel, A. (2021) *Known Unknowns: Data Issues and Military Autonomous Systems,* Report, Geneva: United Nations Institute for Disarmament Research, https://unidir.org/known-unknowns.
105 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', p. 48.
106 In recent years, there has been some progress in experimenting with techniques for vetting AI systems both before and after deployment, but these efforts have highlighted significant hurdles that remain to be solved. For example, a recent expansive red-teaming exercise for a large model developed by OpenAI demonstrated that there was simply no way to anticipate all of the ways the system could be abused, or all of the ways that its biases could result in harm, in advance of those harms actually arising. See Mishkin, P. et al. (2022), 'DALL·E 2 Preview - Risks and Limitations', available on GitHub at https://github.com/openai/dalle-2-preview/blob/main/system-card.md.
107 Boulanin, V. and Verbruggen, M. (2017), 'Article 36 Reviews: Dealing with the challenges posed by emerging technologies', Solna: Stockholm International Peace Research Institute, https://www.sipri.org/publications/2017/other-publications/article-36-reviews-dealing-challenges-posed-emerging-technologies.
108 Brundage et al. (2020), 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', p. 11.

the manner in which they are used.[109] In the absence of major breakthroughs in audit practice and implementation, these will be, as the scholar Mona Sloane has written, 'toothless'.[110] (As with explainability, this is not a reason to abandon audits as a potential solution – but rather a cause to caveat any mention of audits as a fix-all.)

Much of the policy-level language of AI ethics assumes that a computer can replicate the ethical parameters that guide human decision-making.

More fundamentally, much of the policy-level language of AI ethics assumes that a computer can replicate the ethical parameters that guide human decision-making.[111] In reality, human ethical parameters cease to be ethical parameters, as such, when they are translated into the mathematically defined parameters that guide the outputs of the computerized system. There is a significant difference between the mathematically defined processes by which AI systems achieve their goals, for example, and the human capacity to address grey-zone cases, account for uncertainty and engage productively with ambiguity in decision-making.[112] Concepts such as 'fairness' or system 'trustworthiness' therefore cannot be codified with concrete, testable numerical metrics[113] that rate the degree to which an AI system is ethical. At best, machines can only offer an illusion of ethical behaviour – a computational mimicry of ethical decision-making that could fail at the first contact with something unexpected.

Box 2. Moving beyond generalized ethics

If regulatory entities are to develop robust mechanisms to ensure that AI systems are ethical, these mechanisms will need to be highly specific and tailored to each type of AI system, each AI role and every context of use. Measures to ensure that the algorithm guiding a small robotic packing system in a warehouse operates ethically are unlikely to be adequate for a machine-learning system that seeks out instances of welfare fraud or a surveillance tool that identifies 'suspicious behaviour'. A fully tailored framework may need to differentiate systems both in terms of their level of risk *and* their technical architecture.

109 Ng, A. (2021), 'Can Auditing Eliminate Bias from Algorithms?', The Markup, 23 February 2021, https://themarkup.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms.
110 Sloane, M. (2021), 'The Algorithmic Auditing Trap', *One Zero*, 17 March 2021, https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.
111 For example, Singapore's AI strategy describes an effort 'to develop AI technologies that learn like humans, understand humans, and can explain their inner workings to humans', see Singapore Smart Nation and Digital Government Office (2019), 'National AI Strategy', p. 19.
112 Birhane (2021), 'The Impossibility of Automating Ambiguity'.
113 Birhane, A. et al. (2022), 'The Forgotten Margins of AI Ethics', p. 4, https://arxiv.org/pdf/2205.04221.pdf.

Box 3. Investing in ethics

The technical complexity of the challenges described in this section is vast.[114] Therefore, governments that invest more in tools for verification and enforcement as a proportion of total investment in AI may be more likely to succeed. In such regimes, fewer tools that are inappropriate for use will end up being deployed and those tools that are deployed will be less likely to cause unanticipated harms. However, this raises broader questions. Mechanisms for verifying and enforcing this type of AI ethics will be expensive. They will require significant computational resources for processes such as data-driven verification.[115] This is likely to make it hard for low-income states to achieve 'ethical AI' in the mould envisioned by most strategies. And unlike the broader ethical governance measures described below, techno-centric ethical AI tools are unlikely to yield societal benefits beyond realms directly relevant to AI. (Efforts to ground AI ethics in established human rights frameworks[116] are a potentially positive step in this regard.)

A broader challenge

Unless researchers make unprecedented technical breakthroughs on these ethical challenges, any tech-centric vision of AI ethics is likely to 'lack mechanisms to reinforce its own normative claims' as Thilo Hagendorff has written.[117] If anything, as AI systems become more powerful and open-ended, it might get harder to implement and enforce AI ethics by means of the comparatively narrow, tech-focused (and Western-dominated[118]) measures proposed thus far.[119] As such, a comprehensive implementation of 'AI for humanity', as one strategy put it,[120] could require profound structural reforms that go far beyond what has traditionally been considered within the scope of AI ethics.[121]

---

114 Hendrycks, D., Carlini, N., Schulman, J. and Steinhardt, J. (2022), 'Unsolved Problems in ML Safety', https://arxiv.org/pdf/2109.13916.pdf.
115 Brundage et al. (2020), 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', p. 12.
116 Jones, K. (2022), *AI governance and human rights: Resetting the relationship*, Research Paper, London: Royal Institute of International Affairs, https://doi.org/10.55317/9781784135492.
117 Hagendorff, T. (2020), 'The Ethics of AI Ethics: An Evaluation of Guidelines', *Minds and Machines*, 30, pp. 99–120, https://doi.org/10.1007/s11023-020-09517-8.
118 The Global South is conspicuously under-represented in global AI ethics principles. See Fjeld et al. (2020), 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI', p. 7. Non-male authors are also drastically under-represented in the AI ethics literature. See Hagendorff (2020), 'The Ethics of AI Ethics: An Evaluation of Guidelines'.
119 As the 2022 Stanford AI Index Report noted, paraphrasing Jack Clark, one of the project's leads, 'The bigger and more capable an AI system is, the more likely it is to produce outputs that are out of line with our human values'. This correlation between system performance and risks has also been identified by The Future of Life Institute, see Tegmark, M. (undated), 'Benefits & Risks of Artificial Intelligence', Future of Life Institute, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence, as well as Centre for the Study of Existential Risk (undated), 'Risks from Artificial Intelligence', https://www.cser.ac.uk/research/risks-from-artificial-intelligence; and Bender, Gebru, McMilla-Major and Shmitchell (2021), 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'.
120 Villani (2018), *For a Meaningful Artificial Intelligence: Towards a French and European Strategy.*
121 For example, most policy documents do not include measures to address the societal causes of bias, as found in Balayn and Gürses (2021), 'Beyond Debiasing: Regulating AI and its inequalities'.

For example, it is becoming more widely recognized that bias in AI cannot solely be addressed in the system's architecture and data, but rather it must be tackled throughout the development and implementation pipeline.[122] System robustness will need to be addressed not only by improving model performance, but also through improving users' judiciousness in determining whether the system should be deployed in the first place. 'Human control' will not be achieved solely by way of transparent or explainable AI systems – it will rely on rigorous training of the human(s) interacting with a system. Justice in AI will rely on novel human-centric legal instruments to ensure that no harms ever fall within the 'accountability gap',[123] which arises between the technical causes of an AI failure and the human legal agents who must necessarily be held responsible.[124]

When ethical AI is understood in this way, as a sociotechnical paradigm, it is clear that many stakeholders might not be AI-ready. That is, they might not be sufficiently transparent, accountable, or fair to deploy an AI system without causing undue harm. For example, it is not entirely reasonable to expect that a company that built its share value on a first-to-market promise will be capable of implementing safety-critical automated systems that require extensive complex testing and validation frameworks. Institutions embedded with structural inequalities (such as a lack of diversity in leadership positions) cannot necessarily be expected to debias their entire technology development and deployment pipeline if an audit suggests that they should do so.

This issue may be particularly tricky when it comes to organizations whose mode of operating itself stands at odds with the precepts of AI ethics. Non-transparent organizations may struggle to make their AI products fully explainable or accountable. Similarly, one cannot assume that AI-based services provided by organizations whose mission does not advance the cause of fairness (say, for example, a predatory lender) will create equal outcomes for all stakeholders. In the security domain, it is likewise unclear how far AI ethics can be aligned with some of the predominant motivators for AI use – especially increased lethality, increased speed and force multiplication.[125]

This extends to policymakers and regulators themselves. A government whose policies regularly fail to encode fundamental human rights is unlikely to be capable of building ethical AI policies that prevent and mitigate harm equally for all stakeholders. In regulatory environments that lack effective instruments for holding powerful entities accountable, it is by no means guaranteed that

---

122 Schwartz, R. et al. (2022), *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, Special Publication (NIST SP), Gaithersburg: National Institute of Standards and Technology, https://doi.org/10.6028/NIST.SP.1270.
123 Elish, M. C. (2019), 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction', *Engaging Science, Technology, and Society*, 5, pp. 40–60, DOI:10.17351/ests2019.260; Raji, I. D. et al. (2020), 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, https://doi.org/10.1145/3351095.3372873.
124 Put another way, it is not a matter, as some have put it, of 'holding AI accountable' but rather *of holding those responsible for the AI accountable*.
125 The core risks of military AI systems are anticipated to stem directly from the erosion of human control, both in terms of the amount of time that humans have to respond to potential system errors (say, by calling off a weapon) and by the amount of visibility that humans have into the system's processing. See Schwarz, E. (2018), 'The (im)possibility of meaningful human control for lethal autonomous weapon systems', *Humanitarian Law & Policy*, 29 August 2018, https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems.

stakeholders will always be made answerable when the AI for which they are responsible causes harm. States that do not confer equal rights on all citizens – for example, those that criminalize homosexuality or limit citizens' reproductive rights – will be poorly equipped to achieve fairness in public sector AI programmes. States that confer greater rights on wealthy and politically connected members of society cannot be expected to enforce ethical principles in a manner that protects all citizens equally.

Put bluntly, it may simply be impossible to achieve true AI ethics in the absence of broader measures that strive to make the societal context in which AI operates more ethical. This in turn throws some doubt on whether the imperatives of growth, scale and generating returns on investment – imperatives that naturally hold primacy for those seeking 'AI supremacy' – may even be compatible with the precepts of ethical AI. Perhaps the notion of beating one's peers in a race for AI may be diametrically incompatible with the precept of ensuring that AI only ever alleviates existing inequalities and injustices.

A more grounded framing would be to treat fairness and equality and accountability as goals in and of themselves that hold equal, if not greater, importance than relative AI capacity or AI supremacy. Such a framework would treat ethics not as a channel for enabling the highest possible degree of AI uptake, but rather as an opportunity to fashion a society that is fairer and safer for everyone.

06 Recommendations

To recalibrate the AI discourse, we need to shift our perspective, seek new policy assumptions, plan for the worst, and think pre-emptively about whether certain AI applications are actually worth pursuing.

In light of the issues and opportunities described in the preceding chapters, this paper makes the following recommendations.

Recognize assumptions as assumptions. All parties should actively flag when an assumption – as opposed to a ground truth – is being used as the basis for a policy, and provide a framework to consider that assumption's consequences and counterpoint(s). For example, policymakers who recognize that the success of a particular government action hinges on the assumption that growth in the performance of AI systems will continue to be linear in the years ahead could consider the harms or losses that that policy might incur if AI performance flattens. Such an assessment should integrate non-technical perspectives, which could highlight potential externalities that would not be immediately obvious in a strictly technical analysis.

Recognize who these assumptions serve and consider whether they are representative of all stakeholders. As noted throughout this paper, policy assumptions are rarely neutral. They tend to serve a particular set of stakeholders' interests. Identifying the interests embedded in assumptions will make it easier to highlight the political and ideological drivers of the AI discourse, and raise key questions as to whether any resulting policy would be truly representative of the full span of groups that are likely to be impacted by it.

Explore alternative or additional policymaking assumptions. Anticipatory governance relies on assumptions. Therefore, stakeholders should certainly not shy away from making *any* assumptions. Rather, they could seek out grounded, inclusive assumptions that might serve to guide AI policy alongside those assumptions that are already widespread.

Hope for the best but plan for the worst. A genuinely anticipatory style of AI governance anticipates failure and success in equal measure. That is, in addition to seeking the best-case scenario for AI development and implementation, policy

measures should emphasize preventing the worst potential outcomes. When considering a potential role for AI or a potential set of policies, stakeholders should war-game the most detrimental potential outcomes of that policy or AI development and, as needed, include measures to hedge against those outcomes. If acted upon in good faith, such an attitude does not need to stand at odds with the technical community's right to experiment, explore and innovate.

Measure state capacity to adopt AI in a way that truly serves the common good. Metrics of state AI capacity or AI readiness should be expanded to include factors such as openness and transparency of institutions; freedom of civil society and the press; rule of law; economic equality; and educational attainment in other fields outside of science, technology, engineering and mathematics.

Subject AI applications and organizations to *ex ante* audits. Today, there is a strong evidential basis for the claim that some applications of AI simply are not worth pursuing, either because their benefits could never outweigh their harms or because we lack the technical or sociotechnical capacity to ensure that they will be ethical. Yet, as noted above, the discourse often makes little room for a thoroughly anticipatory evaluation of potential risks. Therefore, one way to counteract AI risks is to establish *pre-development* assessments of a proposed system's risks, and to weigh these against its expected benefits.[126] In determining whether to engage in a particular government application of AI or to support the development by a private entity of an AI for a novel application (say through R&D grants), states could develop a process (perhaps executed by an independent body) that exhaustively evaluates:

- The anticipated net benefit of the system. This assessment must be based on a reasonable estimation of technical capacity (that is, an expected benefit cannot be contingent on an as yet unachieved technical breakthrough);
- The risks of the proposed system, both primary (e.g. what would happen if the system fails?) and secondary (e.g. would it require the creation or diffusion of a dataset that is vulnerable to abuse or attack?);
- Whether the entity that is developing and deploying the AI system has the appropriate sociotechnical capability/readiness to create a safe, fair system in a transparent and accountable manner;
- Whether the developing and deploying entity will have the capacity to consistently monitor, respond to, and be held accountable for unanticipated primary and secondary effects;
- Whether the system's ethical problems have known clear solutions or would, instead, rely on unproven technical measures; and
- Whether non-technical or non-AI solutions (which have an existing regulatory infrastructure) could be used in place of the AI system to solve the same challenge with fewer risks.

126 The EU AI draft bill proposes *ex ante* assessments and testing requirements for high-risk AI systems. See 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', p. 14.

Such evaluations would better enable states to predict and avoid risks and harms, and to focus more tightly on proven technologies and solutions. While there are certainly difficulties in anticipating all of the issues that AI might exhibit in real life, this process could be a valuable step in tempering strategies and policies that would otherwise encourage organizations to experiment with the technology as widely and quickly as possible, without regard for either the risks of doing so or the possibility that the system will not actually generate any gains.

About the author

Arthur Holland Michel is a writer and researcher focused on emerging technologies. His work has appeared in *The Economist*, the *Washington Post*, the *Atlantic*, *Wired* and *Vice*, among other magazines and papers. Currently, he serves as a senior fellow at the Carnegie Council for Ethics and International Affairs. Previously, he was a researcher for the United Nations Institute for Disarmament Research and a co-director of the Center for the Study of the Drone at Bard College. His book *Eyes in the Sky*, about the rise of aerial surveillance technology, was published in 2019.

Acknowledgments

This paper is based in part on a virtual expert roundtable held under the Chatham House Rule that took place in March 2022, in collaboration with the Digital Society Initiative. The author wishes to thank all the participants for their valuable contributions during that lively discussion. The author would like to thank Yasmin Afina, Rowan Wilkinson and Marjorie Buchser for their attentive support and guidance, as well as the members of the Chatham House publications team. Finally, he would like to thank the anonymous reviewers for their detailed feedback; the paper is much better for their involvement, whoever they might be.

Cover image: Parallel Universe Exhibition by Ouchhh, Istanbul 24 June 2021.
Photo credit: Copyright © Anadolu Agency/Getty Images

ISBN 978 1 78413 562 1
DOI 10.55317/9781784135621

Cite this paper: Holland Michel, A. (2023), *Recalibrating assumptions on AI: Towards an evidence-based and inclusive AI policy discourse*, Research Paper, London: Royal Institute of International Affairs, https://doi.org/10.55317/9781784135621.

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

Independent thinking since 1920

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