THE UNIVERSITY *of* EDINBURGH
Edinburgh Futures Institute | **Centre for Technomoral Futures**

BRAID
Bridging Responsible AI Divides

Testimony of Prof. Shannon Vallor
Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence
Director, Centre for Technomoral Futures at the Edinburgh Futures Institute
Co-Director, BRAID (Bridging Responsible AI Divides)
The University of Edinburgh

For the U.S. Senate Committee on Homeland Security and Governmental Affairs

Hearing Date: November 8, 2023

Thank you Chairman Peters, Ranking Member Paul, and distinguished Members of the Committee for this opportunity to submit my testimony today. It is a profound honor to address you on a matter of such vital importance to the nation and the human family.

For over a decade, my research as a philosopher of technology has focused on the ethical and political implications of artificial intelligence, robotics and algorithmic automation. I currently direct the Centre for Technomoral Futures at the University of Edinburgh, which integrates technical and moral expertise in new models of responsible innovation and technology governance. I have worked as an AI ethicist in the United States and the United Kingdom, in both academia and the tech industry, and as an independent ethics advisor to UK and Scottish government bodies on public sector uses of AI and data.

In the UK, I lead multi-disciplinary teams of computing, social science and humanities researchers as part of two Responsible AI initiatives funded by the national science funding agency UK Research and Innovation. In the *Trustworthy Autonomous Systems* program, our team is working to strengthen human responsibility for autonomous systems. Our BRAID program, *Bridging Responsible AI Divides*, builds new partnerships between government, voluntary organizations, civil society and industry to drive the growth, embedding and adoption of responsible AI knowledge and practices, while leveraging the arts and humanities to strengthen and sustain the UK's AI ecosystem.

1. Introduction

My research is interdisciplinary, but deeply informed by philosophical and historical perspectives on technology's role in shaping human values, character and capabilities. Among the most important of those capabilities is *self-governance*. This capability, to reason, think and judge for oneself how best to live, underpins the individual civil and political liberties guaranteed by the U.S. Constitution and by international law. It also underpins democratic life. In a democratic society, decisions about how best to live must be made in political cooperation with those whose fates are intermingled with ours.

The link between AI and our capacity for self-governance has two dimensions that interweave in complex ways, not unlike a Mobius strip. One side of the strip is our need, in any democratic society, to jointly exercise responsible self-governance of the new social, political and economic powers that AI technologies inject into our institutions. The second side of the strip is the way that AI technologies can undermine our confidence in, and will to exercise, those same human capabilities of self-governance.

My testimony will expand on these two interrelated challenges for AI and democracy; I will conclude with some brief, but I hope, encouraging reflections on lessons from history for the responsible democratic governance of AI technologies.

2. AI as a Subject of Democratic Governance

As a new source of immense socioeconomic and political power, AI is *something we must govern*. Why? Because ungoverned, or ungovernable, social and political power is deeply incompatible with democratic principles. A core principle of modern democracies is that free peoples may not *justifiably* be subjected to social and political powers which determine their basic liberties and opportunities, but over which they have no say, which they cannot see and freely endorse, and which powers are in no way constrained by or answerable to them.

A point of clarification is needed. AI is not a single technology, but many; AI is not one monolithic power, but a smorgasbord of them. Many types and uses of AI do not affect our fundamental liberties and opportunities at all. So it does not make sense to talk about our need to govern AI *as a whole*. However, for the sake of clarity in this testimony, take 'AI' here to refer to the subset of AI technologies that increasingly *do* affect our basic liberties and opportunities to flourish, either by greatly amplifying the power of certain individuals and existing institutions, or by generating new powers that can be exercised upon us.

Now, in a democratic society, *might does not make right*. Power is not self-justifying. Any exercise of power that extends beyond the private conduct of the individual, in ways that significantly impact the opportunities, welfare and liberties of others, always requires social and political legitimization. We do this through the joint exercise of our political capacity. Legislation and regulation are one way that democratic peoples, through their elected representatives, jointly govern the powers that impact us. Expressing and enforcing shared moral and political norms is another way we collectively self-govern. Adopting professional and technical standards is yet another. Market incentives do some of the work (though far less than some economists imagine), and various forms of public and organizational policy not enacted in law do much of the rest. For most significant forms of power, a combination of these tools is needed. What matters is that the right incentives are created to ensure the responsible, trustworthy and politically legitimate use of that power, under appropriate, mutually agreed upon, and reliably enforced constraints.

In the United States, as with most if not all democratic countries, AI as a new source of power has yet to be socially or politically *legitimized* by the necessary acts of effective governance. These technologies are operating in most domains as ungoverned and growing powers; either because the necessary laws, regulatory bodies, policies, norms and standards to govern them have yet to be created, or more often, because the existing modes of governance that already apply are simply not being used or reliably enforced.

Nor is the growing power that AI technologies generate being distributed equally across our society, so that all of us might choose to use this new power for our benefit. Rather, the power and benefits these technologies afford are increasingly concentrated in very few hands, especially those (corporate) hands that are *already* operating with undue influence on our democratic systems of self-governance. This concentration only amplifies the economic inequality that has been rising in this country for over 40 years, all while intergenerational socioeconomic mobility fell well behind most other developed countries – a linked phenomenon known by economists as the Great Gatsby Curve.[1]

If AI technologies were an equalizing force, as well as transparently and widely beneficial, generating little risk to society or vulnerable groups, then closing the AI governance gap would not be so urgent. Unfortunately, we have a growing pile of evidence for the many harmful effects of AI systems and related technologies on humans living today. While the potential for AI technologies to benefit society remains immense, far fewer of those shared benefits have materialized than were once confidently predicted as imminent.[2] There are many promising applications of AI that we *can* use now, in health research, energy research, and environmental management;[3] but the most socially beneficial use cases suffer from low investment, and very few are being deployed commercially at scale.

[1] Steven N. Durlauf, Andros Kourtellos, Chih Ming Tan (2022), 'The Great Gatsby Curve,' *Annual Review of Economics* 14:1, 571-605. https://www.annualreviews.org/doi/10.1146/annurev-economics-082321-122703

[2] For example, in 2015 IBM's Watson for Health AI was widely predicted to soon outperform human doctors; it crashed out into a commercial and scientific failure just a few short years later. Fully automated, safe and reliable driverless cars, and robots that can relieve exhausted caregivers, have been 'just around the corner' for over a decade.

[3] Sophie Bushwick, '10 Ways AI Was Used For Good This Year,' *Scientific American,* Dec 15 2022. https://www.scientificamerican.com/article/10-ways-ai-was-used-for-good-this-year

The disparity between the timelines of AI benefit and AI risk is striking. The Committee has heard much about AI's risks in previous testimony, and I will not detail them here. Unlike the benefits, the risks have suffered no delay in arriving.[4] And yet we have done vanishingly little about them. Most Americans, and billions of others around the world, remain highly vulnerable and exposed to unjust and discriminatory automated decisions, dangerous AI malfunctions, unwarranted algorithmic surveillance, false arrests and profiling, misidentification, disinformation, fraud, and unsafe or illegal content. As a result, public attitudes toward AI are souring,[5] a serious warning sign for those of us who *want* the technology to mature and succeed. Other technologies, from GMOs to nuclear energy generation, have historically suffered public backlash in ways that greatly limited their beneficial use and advancement, and AI is very much at risk of a similar backlash.

This is not a *scientific* problem. We have studied AI's risks extensively over the past decade, and there is a mountain of evidence about their causes. They are not mysterious, but fairly well understood. And while there is a lot still to learn about how to manage these risks, we are already well on our way. That is the good news! For nearly a decade, researchers in the closely related fields of AI and data ethics, Responsible AI, machine learning fairness, and AI trust and safety have been generating a truly impressive body of powerful tools for documenting, evaluating and auditing AI systems, for anticipating and measuring harmful AI outcomes, and for mitigating their risks.

Organizations like NIST, the IEEE Standards Association, the British Standards Institute, the Alan Turing Institute, the Ada Lovelace Institute, AI Now, the World Economic Forum, and many others have released ample bodies of guidance on how to use these tools. Research programs like BRAID, which I co-direct in the UK, are creating new pathways to embed, test, and adopt Responsible AI tools in industry, government and third sector organizations that want to use AI safely and successfully.[6] If we mandated the judicious and responsible use of these tools tomorrow, across the AI ecosystem, it would take a bit of time for companies to comply and for the effects to show up, but they would come. Sooner, I would wager, than will safe driverless vehicles.

[4] Important advances have been made with the help of AI in areas such as protein-folding research, disease diagnosis, flood forecasting, and automated transcription and captioning tools. But the benefits have thus far been dwarfed by the harmful and unjust patterns that routinely emerge in automated algorithms used in hiring, education, policing, public benefits fraud detection, bail risk, hospital triage, and numerous other high-stakes endeavors. The harms are not theoretical: they include false arrests, lost jobs, bankruptcies, separation of children from innocent families, premature deaths and suicides. See https://incidentdatabase.ai/

[5] Alec Tyson and Emma Kikuchi, 'Growing public concern about the role of AI in everyday life,' Pew Research Center, Aug 28 2023. https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/

[6] https://braiduk.org/fellowships

However, the use of these tools in high-impact and high-risk AI systems remains sporadic, opaque, inconsistent and underincentivized. For example, in 2022 and 2023, just as today's new generative AI tools were being released, several AI companies made heavy cutbacks to their in-house AI ethics and trust and safety teams, or even removed them altogether.[7] Recently, many of the same corporations rebranded such efforts under the label of 'AI safety', with a new focus. Instead of bringing back the experienced, expert teams they laid off, or using the tools those teams created to responsibly address the individual and social harms caused by their existing products and business models, several of these companies now seek government and philanthropic funding to invest in technical study of the unknown, longer-term risks of future 'frontier' models that could be more dangerous than those we have today.[8]

Yet many of the same commercial AI leaders warn that government interference in their efforts, whether through regulatory constraint, demands for transparency, or greater exposure to liabilities for AI harms, will only delay or diminish their capacities to get ahead of these so-called 'existential risks' for our benefit. The implication is clear: as long as we stand aside and let them work (ideally subsidized by public investment), they will keep us all safe from the AI bogeyman that they assure us threatens human eradication.

This problem is not scientific or technical. It is *political*. The history of political thought that shaped this nation tells us how we should assess promises of safety and security from those who seek release from accountability as the payment. As the philosopher John Locke said in 1689 of similar promises from unaccountable monarchs: "This is to think that men are so foolish that they take care to avoid what mischiefs may be done them by polecats or foxes, but are content, nay, think it safety, to be devoured by lions."[9] Nor is safety enough, if it comes at the expense of our liberty and capacity for self-determination. As noted by Jean-Jacques Rousseau in 1762's *The Social Contract,* humans can also find safety and tranquility in dungeons.[10]

I do not blame the powerful corporations behind today's AI technologies for evading their social responsibilities to be accountable and answerable to the rest of us. I have worked with and for technology companies. I have collaborated with many researchers and

[7] Gerrit De Vynck and Will Oremus, 'As AI booms, tech firms are laying off their ethicists,' *The Washington Post,* Mar 30 2023. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/

[8] Parmy Olson, 'There's too much money going to AI doomers,' *The Washington Post,* Aug 16 2023. https://www.washingtonpost.com/business/2023/08/16/ai-apocalypse-there-s-too-much-vc-money-going-to-doomers/ada4388e-3bed-11ee-aefd-40c039a855ba_story.html

[9] John Locke (1689/1884). *Two Treatises on Civil Government.* London: Routledge, p. 230–40.

[10] Jean-Jacques Rousseau (1762). *The Social Contract*, trans. G.D.H. Cole, https://web.archive.org/web/20171215154541id_/https://www.ucc.ie/archive/hdsp/Rousseau_contrat-social.pdf

engineers employed by them (and still do). Virtually all *wanted* to make fairer, more transparent, and more beneficial tools, and most did their best to make that happen, with what power they had. I have spoken at conferences where well-paid machine learning engineers got *teary* telling me how much they wanted to build responsibly. Many of the most groundbreaking and effective research outputs and practical tools in the field of Responsible AI have, in fact, been created by people who were at the time employed by large tech companies, such as Timnit Gebru, Meg Mitchell, and Rumman Chowdhury.[11]

The problem isn't the corporations. The problem is a decades-old political failure to give those companies the proper incentives to align their business activities with the wider public interest, or even to comply with minimal standards of transparency, accountability, safety and fairness. No one questions the need for this compliance in other high-stakes, safety-critical fields, like aviation or pharmaceuticals. Who would suggest (especially in the aftermath of the 737 MAX) that Boeing should have been *freer* from regulatory obligations, and trusted to innovate without oversight, transparency and accountability?

Would anyone suggest that Johnson & Johnson be allowed to release a powerful new heart drug in 'beta', with no independent safety testing and no disclosure of efficacy, risks or side effects beyond what the company elects to disclose? Yet Tesla and Cruise have both enjoyed precisely this license for testing autonomous driving technologies on public roads, at least until the California Department of Motor Vehicles recently suspended Cruise's robotaxi deployment, apparently for concealing safety issues and critical evidence in a case where a pedestrian was struck and dragged under one of the vehicles.[12] Last week, it was revealed by a *New York Times* investigation that these supposedly 'driverless' taxis actually required remote human operators to intervene every few miles of travel – something Cruise had apparently not disclosed in any of its marketing.[13]

Corporations are not natural persons or entities. They are socially constructed and legally empowered entities. They were created centuries ago by government charters and continue to exist by virtue of government and social licenses to operate. They follow the incentives given by the legal and commercial order in which they operate. In this they are no different from anything else that operates without the benefit of a moral and social conscience.

[11] For an excellent overview of their contributions, see *Time* magazine's 2023 list of the '100 Most Influential People in AI': https://time.com/collection/time100-ai/

[12] Russ Mitchell, 'Cruise sidelines entire U.S. robotaxi fleet to focus on rebuilding 'public trust,' *The Los Angeles Times,* Oct 27 2023. https://www.latimes.com/business/story/2023-10-27/cruise-shuts-down-robot-cars-rebuild-public-trust

[13] Tripp Mickle, Cade Metz, and Yiwen Lu, 'G.M.'s Cruise moved fast in the driverless race. It got ugly.' *The New York Times.* Nov 3, 2023. https://www.nytimes.com/2023/11/03/technology/cruise-general-motors-self-driving-cars.html

If you need to change such an entity's behavior, you don't argue with it, or try to shame it. You first study its incentives and then modify them, so that it is sufficiently rewarded for pursuing a better course of action and/or penalized for not doing so.

AI companies and their executives are legally incentivized to maximize shareholder value (too often, in the short term), within whatever constraints the jurisdictions they operate in supply and adequately enforce. Multinational corporations will exploit opportunities for regulatory arbitrage where they are available. But most tech companies will comply with reasonable, well-enforced regulation that raises the floor for responsible conduct, as long as it enables continued commercial viability and innovation (which effective governance does, by greatly lowering the risks for those who invest in, adopt or use the technology).

If a dog chases a kid on a bicycle, I don't blame the dog. I blame the owner who didn't leash it. I don't blame tech companies for not making AI safe, fair and accountable. I blame the democratic political institutions that have neglected their duties to competently govern the powers that now impinge upon their citizens' vulnerabilities and liberties.

It is therefore encouraging to see moves this year in Europe and the UK to consider, albeit in very different ways, empowering regulators to act on AI harms. In the United States, the Blueprint for an AI Bill of Rights and the October 30 Executive Order are welcome steps toward ensuring that AI governance addresses the most urgent risks presented by the technology, while enabling more of its benefits to be developed and distributed to wider publics. It also promises to incentivize more competition in the AI ecosystem, which is another important goal. But the mechanisms of adequate enforcement, and the incentives for corporations to seriously adopt these measures, remain weak and underspecified.

3. AI as a Stressor Upon Democratic Agency: Philosophical and Historical Perspectives

Today, researchers looking at AI's impact on the health of our democracies tend to focus on the impact of disinformation, misinformation and algorithmic echo chambers. As we see in the social media context, these can diminish our political sense of a shared reality, enable new forms of political persuasion and manipulation, and make social cooperation, cohesion and consensus harder for us to build and sustain. These are well studied concerns[14] that I won't testify to further, except to note that some researchers believe that AI 'deepfakes' and other modes of AI disinformation may be less impactful than the lower-cost, lower-effort modes of political disinformation that have been proliferating with little cultural resistance for decades, even before the social media revolution.

[14] Todd C. Helmus (2022). Artificial Intelligence, Deepfakes and Disinformation: A Primer. *The Rand Corporation* https://www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf

What is less often discussed today is the impact of AI technologies on human confidence in our democratic capabilities to reason and govern ourselves, and the downstream effects on our willingness to hold political space for human liberty and self-determination. Yet a slightly longer historical view reveals that philosophers, sociologists, political theorists, and computer scientists have been warning us about this since the start of the digital revolution. 1960s computing pioneer Joseph Weizenbaum anticipated the corrosive effect of algorithmic automation on human liberty, lamenting in 1976's *Computer Power and Human Reason* that just when humans have "ceased to believe in—let alone to trust—[our] own autonomy," we have "begun to rely on autonomous machines."[15]

He was far from the first. In 1954, French sociologist and philosopher Jacques Ellul foresaw an inevitable conflict between human self-determination and machine efficiency, predicting that the efficiency of the technical order (In French, *technique*) would soon be seen as the highest social and moral value. As a consequence, Ellul predicted, human spontaneity and insistence upon our rights of self-determination would be relentlessly pathologized and forced out of the system. He wrote that "The combination of man and technique is a happy one only if man has no responsibility."[16]

That same year, cybernetics pioneer Norbert Weiner, who developed the first theories of machine learning and intelligent automation upon which our modern AI systems depend, warned of a future where humans imagine that they can surrender the burdens and inconvenience of moral and political decision-making to intelligent machines. He wrote that "to throw the problem of responsibility on the machine, whether it can learn or not, is to cast [our] responsibility to the winds, and to find it coming back seated on the whirlwind."[17]

In Isaac Asimov's 1955 short story *Franchise*, he imagined a future of 2008 in which Americans have surrendered our voting power to the supercomputer Multivac, a mechanical 'black box' that in the story's telling, cannot be fully understood even by its own designers, and yet is said to know the collective American political will with even greater precision than Americans do. Today there is no Multivac and most Americans retain the power to vote. Yet we are increasingly being told by powerful AI scientists and business leaders that AI systems will soon be smarter, wiser, and more rational than we are, if they aren't already. They're not just intelligent, we are told, they are becoming *superintelligent*. They aren't just human-like, they're *superhuman*!

[15] Joseph Weizenbaum (1976). *Computer Power and Human Reason: From Judgment to Calculation.* San Francisco: W.H. Freeman & Co., p. 9.

[16] Jacques Ellul (1954/1964 trans.), *The Technological Society.* New York: Vintage Books. p. 136.

[17] Norbert Wiener (1954). *The Human Use of Human Beings: Cybernetics and Society*. Boston: Da Capo Press/Houghton Mifflin, p. 185.

The evidence for such claims requires moving the scientific goalposts for demonstrating ‘intelligence’ and ‘humanness’ by a mile, but in an era of surging political irrationality and division, the implication of such promises is clear: the truly important decisions can’t be entrusted to humans any more. Leave it to the machines, who will exercise the intelligent control and reasonableness that you no longer believe your neighbor has.

OpenAI’s Sam Altman has previously expressed confidence that we are merely the ‘biological bootloader’ for a new, higher form of intelligence that will dwarf ours, not just in brute computational ability (a far more reasonable claim) but in *wisdom.*[18] Despite the mountain of evidence that today’s most powerful AI systems, including GPT models, tend to relentlessly amplify rather than eliminate unfair human bias, Altman has said, ‘I think we’ll find out we can make GPT systems *way less biased* than any human.”[19] Of course, we could strive (as we have for centuries) to become more equitable and fair in our judgments, and GPT, which is trained on our data, could follow our lead. For Altman this is too many steps. He thinks that without our ‘emotional load,’ a mathematical tool like GPT can somehow leapfrog humanity’s stalled moral development, leaving us in the dust.

Turing Award winner Yoshua Bengio says that we may have superhuman AI systems already, since “we call an AI superhuman if it outperforms humans on a vast array of tasks.”[20] Notice the staggering reduction of what it means to be ‘human’ – a human is now simply an *underperforming completer of (computational and economically valuable) tasks*. Indeed, OpenAI has already redefined artificial general intelligence (AGI), which used to mean ‘a machine that can think and reason just like a human,’ as a system that can “surpass human capabilities in a majority of economically valuable tasks.”[21]

None of these AI leaders claim that these systems are sentient. Bengio and Altman do not claim that AI models have emotions, or a moral conscience, or a sense of justice, or the capacity for love, loyalty and self-sacrifice, or aspirations beyond those dictated to them, or the ability to take responsibility for their own actions. These capabilities have now been excised from the concept of general intelligence, since they don’t fit into the box of ‘economically valuable tasks.’ Nor do they apparently remain central, vital capabilities of the human person. *After all, a machine can be ‘superhuman’ without them*.

[18] Sam Altman, ‘The Merge,’ Dec 7 2017. https://blog.samaltman.com/the-merge

[19] Lex Fridman Podcast #367 (2023), ‘Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI,’ https://auphonic.com/media/blog/LexFridmanPodcast367-transcript.html

[20] Yoshua Bengio, ‘FAQ on catastrophic AI risks,’ Jun 24 2023. https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/

[21] Mark Sullivan, ‘Why everyone seems to disagree on how to define AGI,’ *AI Decoded*, Oct 18 2023, https://www.linkedin.com/pulse/why-everyone-seems-disagree-how-define-artificial-general-oaudc/

It is not AI systems that pose the greatest risk to our humanity right now. It's the people telling us that our humane capabilities are *outmoded*, particularly the forms of moral and political agency that are the fragile foundations of democratic life. After all, if these are nonessential for superhuman intelligence, why carry on with them? What place do they have in social decision-making, or our choices about our shared human future? The philosopher and theologian Hans Jonas, like Ellul, Weiner, and Weizenbaum before him, saw this coming. He wrote in 1984, "we need wisdom most when we believe in it least."[22]

Just as when Jonas wrote at the height of the Cold War, today the security of the human family hangs on our capacity for wise self-governance and collective political reason in service of shared values. Yet we believe in these things with far less sincerity and fervor than we did in the 18th century. As philosopher of technology Langdon Winner observed in 1986, even the concept of 'shared values' has lost its capacity to help us wisely steer technological development to desirable ends, since we no longer see 'values' as we once did, as objective features of a desirable and just state of human affairs to be pursued by democratic processes. Instead, we see values as personal, arbitrary whims of individuals, immune to political challenge and reasoned deliberation. We substitute the older, harder question of democracy that this country's founders asked – in Winner's words, 'how are we to live together, gracefully and with justice?'[23] – with the technocrat's cheap replacement: 'how can we calculate the sum of individual human preferences?'

In 1984, Jonas was clear that we must recover our confidence and belief in shared political wisdom if we hope to have any chance to avert irreversible environmental and climate destruction, as well as nuclear and biological holocaust. We can add to that list the need to safely govern AI. We have the same need for democratic renewal now as we did then, only more so, as global confidence in political reason guided by democratic norms is even lower than when Jonas and Winner wrote.

AI in its present form does not promote that renewal. If anything, AI's growing use to automate high-stakes decision-making, alongside the myth of 'superhuman AI,' threatens our already damaged collective confidence in human moral and political agency. AI hype fosters a narrative in which our unique human capacity for practical wisdom – the virtue that Aristotle called *phrónēsis* – is not a political tool that democratic peoples can wield to responsibly craft better futures, but an outmoded relic waiting for the scrap heap.[24]

[22] Hans Jonas (1984), *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago: University of Chicago Press, p. 21

[23] Langdon Winner (1986/2020), *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, Second Edition. Chicago: University of Chicago Press, p. 162.

[24] For more on practical wisdom and AI, see Vallor (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting,* New York: Oxford University Press, and the forthcoming (2024) *The AI Mirror: Reclaiming Our Humanity in an Age of Machine Thinking*, New York: Oxford University Press.

AI is not an intrinsically destructive set of technologies. They aren't even the most dangerous technologies we've ever made (nuclear weapons still hold that dishonor). AI is just coming along at a very bad time, when our political and moral will to exercise democratic wisdom is flagging, and when even our basic belief and confidence in democratic institutions and the value of democratic ways of life are deeply damaged. Surveys of young generations around the world keep telling us this, and AI isn't the cause.[25] Instead, young people cite the failure of their elected political leaders to work together constructively, and reliably serve the common public interest, rather than their donors and most powerful lobbyists.

It may not be the cause of our democratic malaise, but AI, particularly the careless uses and narratives in which AI is painted as a superior replacement for high-stakes human decision-making, is one more force pressing on democratic cultures already riddled with stress fractures. If we don't assert and wisely exercise our shared capacity for democratic governance of AI, *it might be the last chance at democratic governance we get*.

4. Looking backward, to find a road forward

Current media and academic debates about the impact of AI on society rarely take a historical perspective. AI is treated as *sui generis*, a technology unlike anything we have ever encountered, and one which we are repeatedly told that humans are radically unprepared and ill-equipped to govern. This narrative is deeply flawed and misleading, yet it is endlessly amplified by powerful interests who stand to profit from its acceptance.

Like all compelling but incomplete narratives, this one embeds important grains of truth. Today's artificial intelligence technologies *are,* in many important ways, unprecedented. Yet so were the printing press, the steam engine, the airplane and the automobile. You may say: 'Sure, but these tools couldn't destabilize whole societies, or destroy our humanity!' Yet the printing press *was* once widely feared as an agent of widespread political and moral destruction. So was nuclear power. So was genetic engineering. Ironically, even *steam* power sparked 19th century fears about the evolution of intelligent machines that could replace or enslave humans.[26]

[25] Open Society Foundations, 'Generational Shift: New Global Poll Reveals Large Minorities of Young People Lack Faith in Democracy to Deliver on Their Priorities,' Sep 11 2023. https://www.opensocietyfoundations.org/newsroom/generational-shift-new-global-poll-reveals-large-minorities-of-young-people-lack-faith-in-democracy-to-deliver-on-their-priorities; see also Foa, R.S., Klassen, A., Wenger, D., Rand, A. and M. Slade (2020). "Youth and Satisfaction with Democracy: Reversing the Democratic Disconnect?" Cambridge, United Kingdom: Centre for the Future of Democracy. https://www.cam.ac.uk/system/files/youth_and_satisfaction_with_democracy.pdf

[26] See the 'Book of the Machines' in Samuel Butler's 1872 novel *Erewhon* (New York: Penguin).

Not all of these fears were wrong. The printing press *did* radically remake society, although we now think for the better, as it helped to usher in a new political possibility: well-informed peoples capable of democratic self-governance.[27] Nuclear energy and genetic engineering share a more ambivalent destiny. Each has delivered immense benefits, yet they still retain the potential to end intelligent life on this planet.

In comparison with nuclear power and bioengineering, AI has been described as promising even more potential benefits, and equal or greater dangers, including human extinction. Along with many other AI researchers, I believe many of the 'existential risks' of human destruction attributed to AI are wildly exaggerated and scientifically unsound. Those who use the AI risk narrative to draw political and public attention away from the immediate and undeniable existential threats of runaway climate change and growing nuclear destabilization are, in my view, doing serious harm.

Another harmful narrative is that the U.S. must accelerate AI research without regulatory guardrails to avoid losing an AI 'arms race' with China. While much more research is needed to keep up with the new cybersecurity and defense risks presented by AI systems, a runaway proliferation of unsafe, unregulated AI technologies will only endanger us all. New AI techniques are not the sort to remain state secrets for long, and in any case, China has its own internal reasons to keep AI from developing without tight human controls.

Even so, AI *does* present many incontrovertibly real risks to human flourishing, some very grave indeed. For example, even if AI developments do not threaten human survival on their own, AI might be used in ways that undermine our ability to competently manage truly existential climate, nuclear, and bioengineering risks. For that reason I will not use this testimony to press the case against what its critics call AI 'doomerism.'[28]

For if AI *is* a grave existential risk to humanity, the very worst thing we could do is accept the false narrative that AI is a superhuman power beyond our control, something ungovernable. Any competent historical view tells a very different story. The lesson of the 20th century, from nuclear anti-proliferation treaties to voluntary scientific moratoria on germline engineering, to our remarkable successes in strengthening cultures of responsible, safe, and well-regulated innovation in the aviation and pharmaceutical industries, is clear. We have figured out how to govern many technologies that are new, dangerous, opaque, hard to control and contain, and challenging to predict.

[27] This was not possible in earlier eras of 'democratic' government, when 'peoples' could still not rule themselves, but small bodies of wealthy, elite men might vote as equals as they ruled the rest.

[28] Matteo Wong, 'AI doomerism is a decoy,' *The Atlantic,* Jun 2 2023. https://www.theatlantic.com/technology/archive/2023/06/ai-regulation-sam-altman-bill-gates/674278/

We cannot make any new technology perfectly safe and risk-free, AI included. Even well-designed governance systems can fail to prevent harm, or fall victim to regulatory capture. But we absolutely *can* govern AI in ways that enable public trust. We currently lack the political will, but certainly not the capacity, because we've done it before.

Millions of people will board an airplane in this country today without a second thought, knowing that it is statistically far safer than driving. Yet in the 1960s, airplanes fell out of the sky with alarming regularity. Enhancing governance of the aviation sector, by better incentivizing safety investments, mandating global industry cooperation with regulators, and capping liability for responsible actors, *did not halt innovation*. Planes eventually stopped falling out of the sky every week, but they also got more efficient, and faster, with new features and services, while carrying more passengers than ever before.

Of course, AI presents new governance challenges. We cannot use the template of civil aviation regulation for AI; governance is not a copy and paste operation. Yet *every* high-stakes technology has presented unique and novel governance challenges, and until unconditional surrender to powerful corporate interests and lobbies became our political default, we didn't let that stop us.

For example, my PhD student Bhargavi Ganesh has studied the U.S. history of steamboat regulation in relation to the challenges of AI governance, and in April 2023 as part of our BRAID policy program, she presented her early findings to staffers at the UK's Office for AI and Department for Digital, Culture, Media and Sport (DCMS) as a source of lessons for how to successfully govern AI. The U.S. in the 19th century was an ambitious pioneer in successful regulation of a novel, powerful technology like the steam engine, which held huge potential economic benefits but had a perverse habit of violently blowing people up in ways that even technical experts could often not predict or fully explain.

This was not seen as a reason to delay action. Just as with AI today, public fears and distrust of steamboat travel were a serious problem for the industry. Starting in 1838, the U.S. federal government took bold, quick steps, in several stages, regulating flexibly and iteratively, adjusting various industry, operator, and regulatory incentives through successive legislative and professional bodies until the desired results were achieved. These efforts built up the marine safety code and culture that we still rely on today.

It wasn't a one-shot success. The first few attempts didn't create all the right incentives, or empower regulatory authorities to succeed. It took decades for steamboat governance to stabilize and for a marine safety culture to mature. 20th century governance of civil engineering, automobiles, aviation, nuclear power, medical devices, and pharmaceuticals all had their own growing and learning pains, yet each took valuable lessons from the regulatory successes and failures before them. How long before we start with AI? If it might take a decade or more to see the full effect, can we afford to wait? The bills this Committee has passed are an excellent start down this road. The Biden administration's new executive order is similarly welcome, yet limited in its enforcement mechanisms. It will take a concerted and bipartisan effort of Congress to provide them.

What should our priorities be for meaningful and democratic AI governance? Here are three practical recommendations, developed with input from the Ada Lovelace Institute, our partners in the UK BRAID program:

1) **Require AI developers to conduct pre-deployment assessments of potential impact on fundamental rights, with participation from affected groups and users – most urgently for public sector applications of AI for decision support.** This follows recent moves by the Dutch government[29] in the aftermath of the childcare benefits AI scandal which forced the government to resign in 2021, after causing immense suffering for thousands of innocent families and children.[30] Mandating fundamental rights algorithmic impact assessments (FRAIA) would also complement the 2022 AI Workforce Training Act, by enabling those tasked with procuring AI systems for public sector use to get the documentation and evidence they need to make informed decisions. Such a regulatory requirement will require careful attention to ensuring that private companies have adequate incentives and capacity to meaningfully comply, but over time the benefits could be considerable.[31]

2) **Require independent third-party audits for high-risk AI systems throughout their lifecycle.** Algorithmic impact assessments are essential for responsible deployment decisions, but they cannot guarantee that an AI system's safety profile will be as expected once deployed, or stable over time. Downstream interactions or subtle changes in the deployment environment can radically alter a model's performance or outcomes. Post-deployment audits are therefore essential for AI systems that impact people's fundamental rights and opportunities. Third-party audits can be conducted by a regulator, or another entity that follows standardized practices and procedures and is accountable to a regulator

[29] Janneke Gerards, Mirko Tobias Schäfer, Arthur Vankan and Iris Muis, 'Impact Assessment Fundamental Rights and Algorithms,' Mar 31 2022. https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

[30] Melissa Heikkilä, 'Dutch scandal serves as a warning for Europe over risks of using algorithms,' *Politico,* Mar 29 2022. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

[31] Andrew D. Selbst (2021), 'An institutional view of algorithmic impact assessments,' *Harvard Journal of Law and Technology* 35:1, https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

(as in financial services or product safety testing).[32] The U.S. could lead in incentivizing and building out the necessary infrastructure for a third-party audit ecosystem.

3) **Institute stronger mechanisms for contestability, liability, and redress for avoidable and significant AI harms, to re-internalize the costs of preventable harm and developer negligence currently being imposed upon vulnerable publics.** Such mechanisms have played a vital role in building and sustaining the safety cultures of civil aviation, civil engineering and pharmaceutical development. Currently, victims whose fundamental rights are violated by careless AI design and deployment have few paths, if any, to seek recourse. The paths that do exist are often unaffordable or inaccessible to the most vulnerable and impacted groups. New systems of legal liability, contestability and redress are needed to incentivize AI developers and deployers to meet a high standard of care.[33]

5. Conclusion[34]

Some AI leaders, like Meta's Yann LeCun, conclude that because today's AI tools aren't a likely road to AGI, they pose no grave threat to humanity.[35] Unfortunately, these techno-optimists are also mistaken. The threat is there. Many have just misunderstood its nature, because without a philosophical and historical perspective, it can be hard to see that the danger is not really from AI itself.

Had AI arrived at a different historical moment, in a period when global confidence and commitment to democratic norms and values was more robust and secure, I do not think any of the risks we see today would be unmanageable or 'existential' in nature. The extremity of the danger from AI arises from our own current weakened moral and political condition, which has compromised our collective will to legitimize and exercise appropriate control over this new form of power, or direct it to just and beneficial ends.

It is this weakened condition that enables some to see the prospect of automating human culture, judgment and meaning as a *selling point*. Marketers of AI-powered plot generating apps for both children and adults now advertise as a benefit the chance to surrender the most vital parts of storytelling: envisioning where a story might go, what a character's

[32] Deborah Raji's work on this subject leads the field; see most recently Inioluwa Deborah Raji (2022), 'From algorithmic audits to actual accountability: Overcoming practical roadblocks on the path to meaningful audit interventions for AI governance.' In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 5 pages. https://doi.org/10.1145/3514094.3539566

[33] AI accountability and recourse are core priorities of our BRAID research program in the UK; see also this related report from our partner The Ada Lovelace Institute: Ugo Pagallo (2022), 'The way ahead on AI liability issues,' https://www.adalovelaceinstitute.org/blog/the-way-ahead-on-ai-liability/

[34] Elements of this conclusion are adapted from Vallor (2023), published Aug 12 2023 in BBC Science Focus: https://www.sciencefocus.com/future-technology/will-ai-make-humans-dumber

[35] https://twitter.com/ylecun/status/1637603426682150912?s=20

backstory and motivations might be, or what unexpected futures we might open from the present. Millions now use generative AI chatbots to summarize scientific research or government meetings; no need to bother deciding for yourself what novel or important ideas are worth remembering. Your child can use a chatbot to summarize the lecture they didn't attend, and write the exam demonstrating that they understood it. Their professor can then use the very same chatbot to generate the feedback the student receives.

After all, why not liberate ourselves from the work of forming and articulating our own thoughts, telling our own stories, and making our own decisions about what matters? That prospect chilled cybernetics pioneer Norbert Wiener, who saw the power of encoding thoughts into language as the source of our unique cognitive liberty and the heart of our political capacity, "as specifically human as any interest can be. *Speech is the greatest interest and most distinctive achievement of man."*[36] And yet today, large language AI models are built and sold to do the speaking *for us*, as well as the thinking and judging that our power of speech enables (whether we speak audibly or by other signs).

Those who protest handing over our humane capacities for thinking, creating and self-governing to mechanical 'Skinner boxes,'[37] are now often met with the cynical proposal that we ourselves are nothing more than meat Skinner boxes. What more are humans, really, than helpless stimulus-response machines, for whom self-determination is just a comforting illusion?[38]

With AI barely on the horizon, Hans Jonas warned us in 1984 of the existential risk of a future that celebrates the "quenching of future spontaneity in a world of behavioral automata," putting "the whole human enterprise at its mercy."[39] He didn't say whether these automata would be machines or people. I think the ambiguity was intended.

On social media and commercial tech stages, generative-AI evangelists are now asking: what if the future is merely about humans *writing down the questions*, and letting something else come up with the answers? That future is an authoritarian's paradise. Self-governance – not just the ability, but the *desire* and *will* to jointly author our own futures and tell our own stories – is the perpetual enemy of unaccountable power.

[36] Wiener (1954), p. 85.
[37] https://twitter.com/neilturkewitz/status/1662495973438881795?s=20
[38] For a compelling philosophical challenge to this view, see John Martin Fischer, 'Some scientists say we don't have free will. As a philosopher I say, of course we do,' *The Los Angeles Times,* Oct 22 2023. https://www.latimes.com/opinion/story/2023-10-22/humans-free-will-biology-neuroscience
[39] Jonas (1984), p. 118.

Dismantling a democratic way of life is costlier and riskier than convincing people that the rare and hard-won treasure they hold is worthless. In 1911, philosopher and mathematician Alfred North Whitehead claimed that "civilization advances by extending the number of important operations which we can perform without thinking of them."[40] But if you want a future for democratic ways of life, ask yourself: what thoughts do the civilized *keep?*[41]

AI is causing very real harms right now, at scale, from algorithmic discrimination and disinformation, to growing economic inequality, to the surging environmental costs of training computationally intensive models. These urgently demand robust, reasonable, and sustained regulatory and political action to incentivize responsible cultures of safe and trustworthy AI development and use. Without these measures, we will see continued growth of public fear and distrust of AI, and suppressed adoption of its beneficial uses.

But if we fear a future without humanity, or a future without democratic freedoms and meaningful human agency, AI isn't what will steal it from us. The question upon which the future of democracy hangs, and with it our fundamental liberties and capacity to live together and thrive on this planet for very much longer, is not 'what will AI become, and where is it taking us?' That question is only asked by someone who wants you to believe that *you're already out of the drivers seat*.

The real question is the one that the most prescient philosophers of technology and computing pioneers have been asking for 75 years now: what kind of future with AI will we and our elected leaders choose to preserve and sustain, with the power we still retain? One where human autonomy, judgment, and decisions *matter*? Or one where they don't?

Much of the talk about AGI and existential risk is a dangerous distraction from what's going on right in front of us. It's not a violent uprising by machines. It's a slow, quiet *human* devaluation of the political and cultural currency of our own capacity for collective wisdom and self-governance. That's the endgame. Our humanity is the stake.

Thank you to the Chairman, the Ranking Member and the Members of the Committee for the opportunity to testify today.

[40] Alfred North Whitehead. *Introduction to Mathematics*. New York: Henry Holt, 1911, p. 46.

[41] Shannon Vallor (2021), 'The Thoughts the Civilized Keep,' *Noema,* Feb 2 2021. https://www.noemamag.com/the-thoughts-the-civilized-keep/