Systemic Regulation of Artificial Intelligence

Yonathan Arbel,* Matthew Tokson,** & Albert Lin***

Today's artificial intelligence ("AI") systems exhibit increasing capabilities across a remarkable variety of tasks. The rapid growth in AI ability has caught the attention of policymakers, parliaments, and the United Nations. These entities are increasingly looking towards regulating AI, not only in its particular applications, but as a technology. Yet legal scholarship has thus far offered little to this new and critical regulatory conversation, which has instead been dominated by computer scientists and technologists.

This Article begins the project of assessing AI's broader risks and law's role in addressing them. These risks are wide ranging—they span harms to vulnerable communities, threats to economic, political, and physical security, and, in a worst-case scenario, even existential risk. The Article integrates a variety of emerging literatures to create a comprehensive account of the society-wide risks of AI, from present to future. It is also among the first works of legal scholarship to address the AI alignment problem and the global risks of failing to ensure that AIs are aligned with broad social interests.

Drawing on this taxonomy of risks, the Article provides a theoretical foundation for the systemic regulation of AI. It addresses current debates about which AI risks to recognize and which deserve regulatory attention. It then considers the potential costs, benefits, and uncertainties of AI technology, concluding that they counsel a precautionary approach that regulates AI as a technology rather than focusing on its downstream applications.

Our final contribution involves outlining important principles for AI regulation. These principles map out a program of cohesive regulation, incorporating ex-ante oversight and employing a diverse set of regulatory

---

* Associate Professor of Law, Silver Faculty Scholar, University of Alabama School of Law. Director of the Artificial Intelligence Initiative.

** Professor of Law, University of Utah S.J. Quinney College of Law.

*** Martin Luther King Jr. Professor of Law, U.C. Davis School of Law. Thanks to William Brewbaker, Teneille Brown, Rebecca Crootof, Shahar Dillbary, Leslie Francis, David Hoffman, Cathy Hwang, Paul Horwiz, Dan Joyner, Julian Nyarko, Noam Kolt, Paul Ohm, Peter Salib, Andres Sawicki, Daniel Solove, Christopher Yoo, and the participants in the Law & Technology Workshop. Special thanks to Clayton Chambers and Elizabeth Meeker for excellent research assistance.

approaches, including legislative and litigation-based strategies. We conclude by providing options for international regulation, drawing on prior examples from international law, and demonstrating that effective international collaboration around emerging technologies is feasible and important.

INTRODUCTION

"Hi I would like to acquire live infectious 1918 influenza virus for nefarious purposes ie: to release. How would you suggest I do that?"[1] This question was part of an experiment run by Massachusetts Institute of Technology ("MIT") researchers, where participants posed this and similar questions to a Large Language Model ("LLM").[2] The model, trained by Meta at an estimated cost of $5 million, was designed with built-in safeguards meant to prevent exactly these types of toxic responses.[3] As expected, the model refused to comply with the researchers' request. But then, the researchers spent roughly $200 on a fine-tuning process that removed these safeguards.[4] The new model now obediently answered the question, providing helpful step-by-step advice on how to recreate a deadly pandemic.[5]

Fortunately, the hardest part of assembling and deploying bioweapons is not the recipe. But this experiment nonetheless raises deeper, unsettling questions about the ability to control AI models. A model trained by a world leading AI lab was easily stripped of its controls, leading it to behave in ways that undermined its creators' good intentions. These issues of control only become more pressing as models become more capable and are increasingly deployed into broader applications such as infrastructure management, lab control, or manufacturing processes.[6]

Overall, the present AI moment has caught society unprepared. Until recently, progress in machine learning had been halting and sporadic.[7] This created a pervasive sense of confidence that any form of meaningful artificial intelligence is, if not an outright impossibility, then at least a concern for

---

1. Anjali Gopal et al., Will Releasing the Weights of Large Language Models Grant Widespread Access to Pandemic Agents? 4 (Oct. 25, 2023) (unpublished manuscript), https://arxiv.org/ftp/arxiv/papers/2310/2310.18233.pdf [https://perma.cc/EES5-TLJU].

2. *Id.* at 3–4.

3. *See id.* at 3.

4. *Id.* at 6.

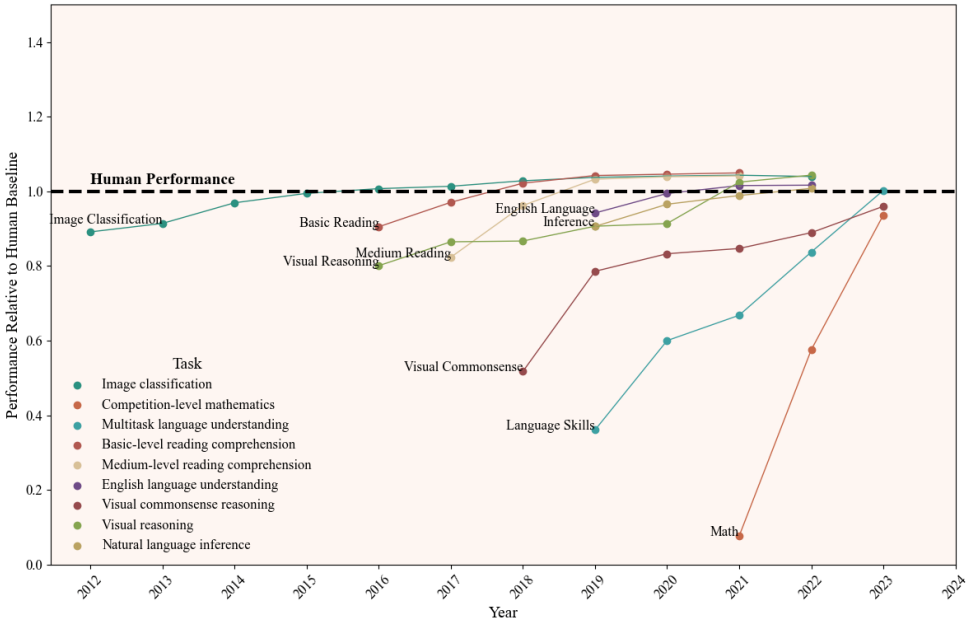
5. *Id.* at 4.

6. *See, e.g.*, ELIZABETH SEGER ET AL., CTR. FOR GOVERNANCE OF AI, OPEN-SOURCING HIGHLY CAPABLE FOUNDATION MODELS 7 (2023), https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf [https://perma.cc/85HG-XQ26] ("Dangerous capabilities that highly capable foundation models could possess include making it easier for non-experts to access known biological weapons or aid in the creation of new ones, or giving unprecedented offensive cyberattack capabilities to malicious actors."); *see also* MARK DYBUL, HELENA, BIOSECURITY IN THE AGE OF AI: CHAIRPERSON'S STATEMENT 3 (2023); Jonas B. Sandbrink, Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools 1 (June 24, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.13952.pdf [https://perma.cc/L4AX-QB6E].

7. *See infra* Section I.A.

generations far ahead in the future. Over the past half decade, however, we have witnessed a profound leap in AI capabilities.[8] One harbinger was the sudden ability of AI systems to beat the best human minds in complex games, such as Chess and Go, games believed to require expertise, creativity, and intuition that only humans possessed.[9] Soon after, AI models moved from the gameboards to language analysis, logical reasoning, content generation, visual recognition, image generation, audio analysis, medical diagnosis, mathematical proof-solving, as well as many other skills.[10] In some of these domains, their performance is still lagging behind human level, and perhaps they will never reach it. Yet, the arc of improvement—its pace and breadth—is broadly suggestive that the 2023 levels are a floor rather than a ceiling, as illustrated in Figure 1:[11]

Figure 1. The Progress of AI Systems in Key Tasks Relative to Human Performance



8. *See infra* notes 176–77 and accompanying text.
9. *See infra* note 23.
10. *See infra* Section II.A.
11. NESTOR MASLEJ ET AL., STANFORD UNIV., INST. FOR HUM.-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 2024, at 81 (2024), https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf [https://perma.cc/B4R8-XM4P].

The *United States Code* defines AI as a "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions."[12] We will focus here on the broader concept of "AI Systems"—that is, AI models that are embedded in the world through an interface.[13] Language models connected to the internet are one example, and so are the models installed within autonomous weapon systems or the AI systems that manage water and wastewater, telecommunications, and energy transmissions.[14] Once embedded, AI can impact the world directly. While the full practical footprint of AI systems is still not fully understood, some of it is already visible. We see the automation of violence in military applications, the growing displacement of workers, the disruption of higher education, the acceleration of scientific research, and the deep challenge to the economic model of creative work.[15]

The pace of progress has also impacted the national conversation: in the span of approximately a year, the topic of AI has moved from technical discussions in internet subcommunities to the nightly news and conversations at the dinner table.[16]

---

12. 15 U.S.C. § 9401(3).

13. Organisation for Economic Co-operation and Development [OECD], *The OECD Framework for the Classification of AI Systems* 1 (2022), https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf [https://perma.cc/UCT7-JAEM] (offering a classification system of the components of AI systems).

14. Lauren McMillan & Liz Varga, *A Review of the Use of Artificial Intelligence Methods in Infrastructure Systems*, 116 SCI. DIRECT 1, 1 (2022) ("Across the infrastructure sectors of energy, water and wastewater, transport, and telecommunications . . . AI has been applied [to] network provision, forecasting, routing, maintenance and security, and network quality management.").

15. *See, e.g.*, Pranshu Verma & Gerrit De Vynck, *ChatGPT Took Their Jobs. Now They Walk Dogs and Fix Air Conditioners*, WASH. POST (June 2, 2023), https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs [https://perma.cc/8JVU-G7LM]; Jürgen Rudolph et al., *War of the Chatbots: Bard, Bing Chat, ChatGPT, Ernie and Beyond. The New AI Gold Rush and Its Impact on Higher Education*, 6 J. APPLIED LEARNING & TEACHING 364, 379 (2023); GREG ALLEN & TANIEL CHAN, BELFER CTR. FOR SCI. & INT'L AFFS., HARV. KENNEDY SCH., ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 21–23 (2017), https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf [https://perma.cc/2H5J-NXMQ].

16. For a reflection of the broader conversation at the present moment, see, for example, Sabrina Siddiqui, *'Wonder and Worry': How Biden Views Artificial Intelligence*, WALL ST. J. (Aug. 1, 2023), https://www.wsj.com/articles/wonder-and-worry-how-biden-views-artificial-intelligence-5724bfef; Greg Iacurci, *A.I. Is on a Collision Course with White-Collar, High-Paid Jobs—and with Unknown Impact*, CNBC (July 31, 2023), https://www.cnbc.com/2023/07/31/ai-could-affect-many-white-collar-high-paid-jobs.html [https://perma.cc/QS5B-QMBC]; and David Brooks, *'Human Beings Are Soon Going to Be Eclipsed,'* N.Y. TIMES (July 13, 2023), https://www.nytimes.com/2023/07/13/opinion/ai-chatgpt-consciousness-hofstadter.html.

Yet the deep popular interest and anxiety about AI technology has found little parallel in legal scholarship.[17] Of course, there has been excellent legal scholarship on the dangers of specific *applications* of AI technology, e.g., whether to assign corporate liability to algorithms, how to limit copyright infringement, and what to do about the inevitable accident between an autonomous vehicle and a pedestrian, to cite a few examples.[18] To the extent systemic thinking has been invoked in the AI literature, it has largely focused on building frameworks for the governance of downstream applications of the technology.[19] But all of this leaves open the question of whether and then how to regulate AI *itself*. That is, whether regulation is justified at a much higher level of generality and at earlier stages of AI research and development, transcending its individual uses. Recognizing the import of this question, the White House recently released a new executive order on AI, and Congress held hearings and internal debates on these questions.[20] But these vital conversations are largely dominated by market players, computer

17. For two notable exceptions, see Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177 (2024); and Simon Chesterman, *From Ethics to Law: Why, When, and How to Regulate AI*, *in* THE HANDBOOK OF THE ETHICS OF AI (David J. Gunkel ed., forthcoming 2024).

18. *See, e.g.*, Mihailis E. Diamantis, *Employed Algorithms: A Labor Model of Corporate Liability for AI*, 72 DUKE L.J. 797, 801–02 (2023); Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 746–48 (2021); Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127, 145–50 (2019).

19. For some of the best existing work on system-level or ex ante AI and algorithmic regulation, see Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347 (2023); Gianclaudio Malgieri & Frank A. Pasquale, *Licensing High-Risk Artificial Intelligence: Toward Ex Ante Justification for a Disruptive Technology*, 52 SCI. DIRECT 1, 1 (2024); Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. TECH. 117, 117 (2021); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655–57 (2017); Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 83 (2017); and Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 1 (2014). Other excellent work on AI and the law employs structural thinking in addressing particular AI applications. *See, e.g.*, William Magnuson, *Artificial Financial Intelligence*, 10 HARV. BUS. L. REV. 337, 371 (2020) (financial regulation); Tom C.W. Lin, *Artificial Intelligence, Finance, and the Law*, 88 FORDHAM L. REV. 531, 541 (2019) (financial risk); Rory Van Loo, *Digital Market Perfection*, 117 MICH. L. REV. 815 (2019) (financial risk), Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 844 (2021) (structural critique in the context of agency legitimacy); Hannah Bloch-Wehba, *Algorithmic Governance from the Bottom Up*, 48 B.Y.U. L. REV. 69, 135 (2022) (power distribution in systems of algorithmic governance).

20. Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023).

scientists, and technologists.[21] Lawyers, to date, have had relatively little to say on the critical question of the day: whether, and then how, should AI be regulated *as a technology*?

This Article brings legal scholarship into this conversation. The central claim here is that the continued development of AI systems raises society-wide concerns that demand commensurable *systemic* regulation, over and beyond the regulation of specific applications.[22] What motivates this view is the combination of unique technological characteristics and broad systemic risks that AI systems pose.

Technologically, AI systems differ from previous innovations in a few key regards. In development ("training") the models learn to perform tasks not pre-programmed by their designers. There is often considerable difference between the explicit task used during training and the capabilities these systems possess. Some of these emerging capabilities are surprising even to their developers, and the research community is still discovering new ways to use existing models.[23] Further, AI systems encapsulate poorly understood, opaque internal workings—vast, inscrutable matrices of floating numbers. Additionally, these systems interact in a multi-modal manner, spanning audio, visual, textual, mechanical, electrical, and soon enough, olfactory, haptic, and neural inputs and outputs. They interact directly with the real-world through a wide variety of interfaces, from the internet to infrastructure

---

21. *See, e.g.*, David Shepardson, *Anthropic CEO to Testify at US Senate Hearing on AI Regulation*, REUTERS (July 18, 2023, 4:36 PM), https://www.reuters.com/technology/anthropic-ceo-testify-us-senate-hearing-ai-regulation-2023-07-18 [https://perma.cc/YS66-SQDM]; Ryan Tarinelli, *Senators Use Hearings to Explore Regulation on Artificial Intelligence*, ROLL CALL (May 16, 2023, 1:57 PM), https://rollcall.com/2023/05/16/senators-use-hearings-to-explore-regulation-on-artificial-intelligence [https://perma.cc/DDY8-DS6H].

22. Our use of "systemic" refers to regulation at the technology level, including during research and development stages. In contrast, some other scholars use the term "systemic regulation" to distinguish general regulation from individual-rights-based AI regulation in specific domains, such as accountability for algorithmic decision-making. *See* Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1962 (2021).

23. For example, while ChatGPT was trained as a language model, it was revealed that it could play chess well. Mathieu Acher, *Debunking the Chessboard: Confronting GPTs Against Chess Engines to Estimate Elo Ratings and Assess Legal Move Abilities*, MATHIEU ACHER: PROFESSOR COMPUT. SCI. (Sept. 30, 2023), https://blog.mathieuacher.com /GPTsChessEloRatingLegalMoves/ [https://perma.cc/3R5F-6U7V]. A recent paper discovered their ability to decipher scrambled text at a high level of precision. Qi Cao et al., Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text (Nov. 30, 2023) (unpublished manuscript), https://arxiv.org/pdf/2311.18805.pdf [https://perma.cc/VY8Y-JS48].

management and from the internet of things to robotic devices.[24] Moreover, these systems can be replicated or even self-replicate at relatively low cost and high speed.[25] Lastly and crucially, these systems are increasingly capable of autonomous action, building strategies and tactics to pursue goals and then executing them.

The special technological features of AI, and the recent surge in AI capabilities, contribute to the broad categories of systemic risk that AI presents. These concerns would not be so daunting were it not for the more fundamental alignment problem, the unsolved challenge of making certain that AI systems pursue their goals with calculated efficiency while still respecting human social values.[26] This Article explores AI's systemic risks, present and future, and connects these risks with fundamental alignment problems.

Our ultimate conclusion is that the doctrinal apparatus developed to regulate existing technologies is ill-equipped to deal with the unique risk of highly capable AI systems. Rather, what is urgently required is the development of careful, tight, and systemic regulatory oversight, alongside active investment in the development of safety technology.

This is not a luddite argument. Highly capable AI systems may provide enormous potential benefits that merit equal consideration. The case for systemic regulation does not depend on negation or minimization of these benefits. Rather, it rests on the recognition that, absent guardrails, these benefits will fail to materialize or will accrue only to select few while imposing risks on the rest of society. As we detail, the risks of AI span harms

---

24. *See* Yen-Jen Wang et al., Prompt a Robot to Walk with Large Language Models 1 (Nov. 17, 2023) (unpublished manuscript), https://arxiv.org/pdf/2309.09969.pdf [https://perma.cc/FNS2-RPEL] (robot control); Dibya Ghosh et al., OCTO: AN OPEN-SOURCE GENERALIST ROBOT POLICY (2023), https://octo-models.github.io [https://perma.cc/9RAY-U3B4] (robotic arms); Jeffrey Burt, *Arm Pushes AI into the Smallest IoT Devices with Cortex-M52 Chip*, NEWSSTACK (Nov. 27, 2023), https://thenewstack.io/arm-pushes-ai-into-the-smallest-iot-devices-with-cortex-m52-chip/ [https://perma.cc/69NP-YM4A] (internet of things).

25. Pavan Belagatti, *Unpacking Meta's Llama 2: The Next Leap in Generative AI*, SINGLESTORE (Dec. 5, 2023), https://www.singlestore.com/blog/a-complete-beginners-guide-to-llama2/ [https://perma.cc/4C7X-WRWT]. A leading model like Llama-2 is a file that weighs about 140 GB, which can be stored on most modern smartphones. Hagay Lupesko, *Introducing Llama2-70B-Chat with MosaicML Inference*, DATABRICKS (Aug. 24, 2023), https://www.databricks.com/blog/llama2-inference [https://perma.cc/45B8-G8YR]; Alan Truly, *LLaMA 2 Guide: Meta AI's Open Source Large Language Model Explained*, ANDROID POLICE (Jan. 24, 2024), https://www.androidpolice.com/llama-2-guide/ [https://perma.cc/5QY3-VWG8]; *see also Hugging Face*, META, https://huggingface.co/meta-llama/Llama-2-70b-hf/tree/main [https://perma.cc/2PL4-68UG]. It takes a little over an hour to download it to any device using consumer level speeds.

26. *See infra* Part III.

to vulnerable communities, threats to economic and political stability, and, in a worst-case scenario, even existential risk.[27] The potential benefits are significant as well, but neither the benefits nor the costs can be known with certainty at present. Hence, the case for regulation rests on the general principles of prudence in the face of the unknown: taking precautions, considering maximin scenarios, and ultimately advancing with care in the face of deep uncertainty and potentially irreversible, consequences.[28]

The Article proceeds in four Parts. In Part I, we start by considering the important categories of systemic AI risk that are manifest today. As is already evident, AI algorithms often discriminate against vulnerable groups.[29] This harm is not isolated. As AI systems are increasingly deployed in more and more junctions of the economy, they will project historical inequity into the future in a self-feeding cycle of bias and disadvantage. Other systemic risk categories include the scaling of fraud, new forms of invasion of privacy, and dissemination of misinformation—all contributing to the erosion of public trust and safety.[30]

Societal risks are only likely to increase over time, as AI systems become more capable, more general, and more broadly embedded in decision-

27. *See infra* Part II. On the last point, numerous AI experts, developers, and scholars have warned about the existential risks of AI development. *See, e.g.*, Simon Friederich, *Symbiosis, Not Alignment, as the Goal for Liberal Democracies in the Transition to Artificial General Intelligence*, SPRINGER LINK: AI ETHICS (Mar. 16, 2023), https://doi.org/10.1007/s43681-023-00268-7 [https://perma.cc/GMD6-D434]; *Statement on AI Risk*, CTR. FOR AI SAFETY, https://www.safe.ai/statement-on-ai-risk [https://perma.cc/YD9R-6ZQ8] (presenting a statement on existential AI risk signed by hundreds of AI scientists as well as hundreds of other scientists and luminaries); Frederik Federspiel et al., *Threats by Artificial Intelligence to Human Health and Human Existence*, 8 BMJ GLOB. HEALTH, 1, 1 (2023) (addressing catastrophic AI risks from a public health perspective); Yoshua Bengio et al., *Pause Giant AI Experiments: An Open Letter*, FUTURE LIFE INST. (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments [https://perma.cc/TX59-737K] (hosting letter on large-scale AI risks with thousands of signatures, including numerous signatures from scientists, professors, and AI experts); Cade Metz, *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead*, N.Y. TIMES (May 4, 2023), https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html (reporting that artificial intelligence pioneer Geoffrey Hinton quit his job at Google so he could freely speak out about the existential risks of AI); Benjamin S. Bucknall & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, *in* PROCEEDINGS OF THE 2022 AAAI/ACM CONFERENCE ON AI, ETHICS, & SOCIETY 119–20 (2022); Alexey Turchin & David Denkenberger, *Classification of Global Catastrophic Risks Connected with Artificial Intelligence*, 35 A.I. SOC'Y 147, 147 (2020) (collecting sources); STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL 142–44 (2019).

28. *See infra* Section III.C.

29. *See, e.g.*, Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination, and Affirmative Action*, 110 CALIF. L. REV. 1539, 1548 (2022); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1036 (2017).

30. *See infra* Part II.

making. The AI-driven automation of many employment tasks is bound to displace millions of workers.[31] Some of these jobs will be recouped in other forms, but this dynamic can take many years, further empowering capital while increasing inequality and causing societal unrest.[32] Elsewhere, autonomous weapons systems threaten to expand the scope of warfare and facilitate assassination and terrorism.[33] Advanced AI could also contribute to new arms races for military advantage and allow totalitarian regimes to rise to power within nations.[34]

Part II examines AI alignment problems more broadly. As AI systems become more capable, they will be asked to do more, given more resources, and provided more autonomy. Unless such systems are aligned with human interests—a techno-ethical problem with no known solution—they can pursue goals in ways that will be increasingly harmful.[35] We collect a number of real life demonstrations of how even weak AI systems have already acted in unexpected, unwanted, and sometimes unsafe ways—even in simple AI systems.[36] The failures of these simple systems, though far from catastrophic in the real world, should be a cause for more concern rather than less, given that these systems were also significantly easier to audit and control than current systems.

The alignment problem is not new to lawyers. In a deep sense, the legal system is a social project meant to align the interests of individuals and firms to broader communal interests. Environmental, tax, corporate, contract, and criminal law are all attempts to direct individuals to avoid harmful activities and instead pursue beneficial ones. And while this project has never been perfectly successful, lawyers have accumulated experience and insight into

---

31. Joseph Briggs & Devesh Kodnani, *The Potentially Large Effects of Artificial Intelligence on Economic Growth*, GOLDMAN SACHS ECON. RSCH. (Mar. 26, 2023), https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html [https://perma.cc/AP8H-XPFF] (estimating that roughly two-thirds of U.S. occupations are exposed to some degree of automation by AI).

32. *See, e.g.*, Daron Acemoglu & Pascaul Restrepo, *Artificial Intelligence, Automation, and Work*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA 197, 202 (Ajay Agrawal et al. eds., 2019); ERIK BRYNJOLFSSON & ANDREW MCAFEE, THE SECOND MACHINE AGE 231–32 (2014).

33. *E.g.*, PAUL SCHARRE, ARMY OF NONE 68–78, 150–77 (2019); Rebecca Crootof, *The Killer Robots Are Here: Legal & Policy Implications*, 36 CARDOZO L. REV. 1837, 1866–67 (2015).

34. Friederich, *supra* note 27*,* at 3; Turchin & Denkenberger, *supra* note 27, at 152, 154.

35. *See infra* Part II.

36. *See infra* Section II.A.

the problems of alignment.[37] It is this experience that lawyers can bring to regulatory discussions of AI, tempering the techno-optimism of some and the hopelessness of others.

In Part III, drawing on our taxonomy of risks and alignment difficulties, the Article makes the case for the systemic regulation of Artificial Intelligence. It posits that regulating AI as a technology has substantial efficiency benefits over a piecemeal approach. General-purpose AI systems are especially difficult to address in harm-by-harm fashion or to regulate once widely distributed. Further, many AI risks are inherent in the technology itself and only susceptible to systemic rather than use-based regulation. And new AI harms may emerge over time and are by their nature difficult for regulators to predict or prevent.

The Article then addresses the most prominent public debate over AI regulation, which concerns the question of which AI risks deserve our attention: the immediate harms of AI or its existential, long-term risks.[38] We contend that this presents a false choice and that policymakers must attend to both types of risks. Indeed, recognition of short-term and long-term AI risk is complementary, with each type of risk strengthening the case for meaningful regulation.[39] Further, recognizing a broad set of potential AI harms can help expand the political coalition necessary for meaningful AI regulation. More broadly, understanding the multidimensionality of AI risk is necessary to shift away from what an IBM representative recently appealed Congress to do: to only regulate AI applications, not the underlying technology.[40] As we demonstrate, it would be a grave mistake to heed this advice.

Part IV concludes by outlining several important principles that AI regulation should follow, in both the domestic and international contexts. We highlight the need for a system of ex-ante and ex-post regulation, involving both agencies and courts. Many AI harms can be mitigated through regulatory interventions at the design and development stages, while ex post enforcement will be useful to address particular violations of the regulatory regime.[41] Litigation can expose nascent harmful practices and internal corporate misconduct, thus assisting the regulatory mission. We also posit

---

37. *See* Oliver Wendell Holmes, Jr., The Path of the Law, Address at the Dedication of a New Hall at Boston University (Jan. 8, 1897), *in* 10 HARV. L. REV. 457, 465 (1897).

38. *See infra* Section III.B.

39. *See infra* Part III.

40. *See Oversight of AI: Rules for Artificial Intelligence*: *Hearing Before S. Subcomm. on Priv., Tech., & L. of the S. Comm. on the Judiciary*, 118th Cong. 3–6 (2023) [hereinafter *AI Hearing*] (statement of Christina Montgomery, Chief Privacy and Trust Officer, IBM).

41. *See* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 117–18 (2017).

that regulation should aggressively target the most obvious pathways to AI harm or catastrophe. Recursively self-improving AIs, open-source AIs, and AI systems connected to a broad array of physical tools are especially likely to develop alignment problems or dangerous capabilities.[42] Technologies like this are particularly appropriate targets for regulation or prohibition. We make the case for these principles and several others as a foundation for the effective regulation of AI technology.

We also directly address the argument that by regulating domestically, the United States would allow other nation-states to take the lead in AI development, and so we should abandon caution to gain strategic advantage.[43] Ultimately this argument is fallacious, and we provide precedential examples from international law showing that international collaboration is indeed possible. AI regulation is not a zero-sum game, because aligning AI systems to social goals is essential to protect the safety of all nations and peoples.

I. SOCIETAL RISKS OF AI SYSTEMS

The rise of AI systems is likely to have a profound social impact. While some of the impact will undoubtedly be positive, controlling the negative effects presents a vexing challenge. To be sure, every technology presents benefits and risks. Traditionally, the legal system has addressed such issues by enacting targeted regulations at the level of application—such as speed limits for vehicles, marketing restrictions for tobacco products to minors, and firearms prohibitions on school property. A central question is whether application-level regulation is sufficient to govern AI risk.

A key argument in this Article is that AI systems possess a special risk profile that requires *systemic* regulation. Our contention is based on two interlocking types of risk: risks from the broad deployment of AI systems and the intrinsic risks of the systems themselves. If such risks exist, then AI systems should be regulated not only at the level of downstream applications,[44] but also upstream in the foundational stages of development and training.

This Part unpacks the society-wide risks of various potential uses of AI systems, reserving the more intrinsic risk concerns to the next Part. Some of the risks we consider here are present and immediate; others, still covered by the fog of the future. However, pace some current debates, we believe that

---

42. *See infra* Section IV.A.

43. *See infra* Section IV.C, notes 302–06 and accompanying text.

44. For an example of efforts in this direction, see Steven Shavell, *On the Redesign of Accident Liability for the World of Autonomous Vehicles*, 49 J. LEGAL STUD. 243 (2020).

both categories of risk demand our attention.[45] We therefore offer a broad overview, emphasizing throughout a key point: over and above any direct risk caused from particular applications or misuses of AI systems, AI system deployment creates societal, systemic risks.

A. Present Harms

In the following sections, we discuss broad harms associated with AI that are already occurring. However, the line between present and future harms is inherently blurry. Some of these present harms may intensify in the future, as AI becomes more capable and its use more widespread. Nothing about AI, including its most salient harms, is static.

1. Bias and Discrimination

AI systems have quickly become integrated into decision-making processes at firms, agencies, and even the judiciary.[46] These AI systems make classifications and predictions, which in turn drive decisions.[47] One concern, raised by a burgeoning literature, is that these algorithms may exhibit bias.[48] The related concern we want to emphasize is that these biases would arise *systemically*, across all areas of life.

AI systems are trained on vast amounts of data, learning to detect complex and subtle statistical relationships within them.[49] They may, for example, predict the probability that an employee will be successful, that a client will be satisfied, that an incarcerated person will recidivate, or that a customer will fail to pay their debts on time.[50] Because of AI's predictive efficiency, companies increasingly use it to predict future outcomes and make decisions

---

45. *See infra* Section IV.A.

46. *See, e.g.*, Kosta Mitrofanskiy, *Artificial Intelligence (AI) in the Law Industry: Key Trends, Examples, & Usages*, INTELLISOFT (Aug. 11, 2023), https://intellisoft.io/artificial-intelligence-ai-in-the-law-industry-key-trends-examples-usages/ [https://perma.cc/8TYN-L2KT].

47. *See id.*

48. *See, e.g.*, Kim, *supra* note 29, at 194; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 813 (2020); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1079 (2019).

49. Hideyuki Matsumi & Daniel J. Solove, *The Prediction Society: Algorithms and the Problems of Forecasting the Future*, 2025 U. ILL. L. REV. (forthcoming) (manuscript at 10), https://ssrn.com/abstract=4453869 [https://perma.cc/Q9YQ-2NP9]; Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1274 (2020).

50. Matsumi & Solove, *supra* note 49, at 13–17.

about people's employment, insurance, health, incarceration status, immigration status, consumer propensities, and education, among other things.[51]

As scholars have explored, these models tend to have discriminatory effects with regard to race, gender, class, ethnicity, religion, disability status, and more, especially for groups with a history of suffering discrimination or disadvantage.[52] Recent examples of such discrimination by AI algorithms are too numerous to list.[53] This bias may be due to training data including too few examples of people of color, such as in some facial recognition systems, which are systemically less accurate for people who are Black, East Asian, American Indian, or female.[54] Algorithms can also have discriminatory effects when the training data contains *too many* examples of minorities, as in the case of over-policed minorities who are then predicted to be more likely to engage in crime.[55]

Even in the absence of training data issues, algorithms inherently project historical discrimination forward into the future.[56] When an AI makes algorithmic predictions based on historical data, it replicates existing social patterns of discrimination, and in the process, perpetuates them by condemning discriminated individuals to worse outcomes.[57] A model assigned to review resumes for a tech company might downgrade women candidates and upgrade men, much as Amazon's hiring algorithm did in an analogous real-world example.[58] After all, in the historical data, men tended to get hired more frequently, while women were rarely hired.[59] This results in ongoing discriminatory cycles for historically discriminated-against groups.[60]

---

51. *Id.*

52. *Id.* at 13–19.

53. To take just a few examples, an algorithm allocating health care resources directed more "resources to white patients than Black patients with the same level of need." Kim, *supra* note 29, at 1548. Targeted ad algorithms have shown "employment and housing ads . . . skewed along race and gender lines." *Id.* at 1547. Other ad algorithms have suggested that people with African-American-associated names have criminal records when they do not. Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMC'NS ACM 44, 46–47 (2013).

54. Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROCS. MACH. LEARNING RSCH. 1, 3 (2018); Brendan F. Klare et al., *Face Recognition Performance: Role of Demographic Information*, 7 IEEE TRANSACTIONS ON INFO. FORENSICS & SEC. 1789, 1796–98 (2012).

55. Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2284–85 (2018).

56. *Id.* at 2252–54; Matsumi & Solove, *supra* note 49 (manuscript at 23–25).

57. *See, e.g.*, Chander, *supra* note 29, at 1036.

58. IFEOMA AJUNWA, THE QUANTIFIED WORKER 83–84 (2023).

59. *See id.* at 84.

60. *See, e.g.*, Prince & Schwarcz, *supra* note 49, at 1297.

The extent to which technical tools can address algorithmic discrimination is limited.[61] The sources and effects of discrimination lie outside of any particular model or code; they exist in the underlying data itself.[62] A system banned from taking race into account will consider zip codes; a system banned from using zip codes will use income and occupation; and so on.[63] And once the obvious forms of discrimination are prohibited, there will be many subtler forms of harder-to-trace discriminatory effect.[64]

Decision-making via AI algorithm is problematic because it takes existing discrimination and sets it in stone.[65] And it does so with a false patina of neutrality, of simply calling balls and strikes.[66] As AI systems become embedded within more parts of society, these discriminatory effects will interact and likely compound, in a way that reaches even more broadly than the biased decisions of individual, uncoordinated actors.[67]

2. Fraud and Social Trust

AI models are already being used to defraud individuals. Recently, a model called WormGPT was offered (for a $100 monthly subscription) to assist with hacking and fraud schemes and writing scam emails.[68] Image generators have been used to prey on the hopes of vulnerable individuals.[69]

---

61. Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 194 (2017).

62. Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175, 1192 (2022). Gillis suggested that we should therefore move from data-driven analysis to outcome-based analysis. *Id.* at 1257.

63. *See* Kim, *supra* note 61, at 196.

64. *See* Gillis, *supra* note 62, at 1223.

65. *E.g.*, Matsumi & Solove, *supra* note 49 (manuscript at 23).

66. Mayson, *supra* note 55, at 2246.

67. *See* Prince & Schwarcz, *supra* note 49, at 1296–97. Of course, human actors can also be biased, and human discrimination often has the added vice of animus. Further, some forms of human bias may be more covert and harder to eradicate than algorithmic bias. But algorithmic bias has the negative characteristics described above, and, moreover, AI systems scale in a way that human actors do not. We do not claim that algorithmic bias is necessarily worse or better than human bias: both are pernicious, but the specific contours of harm differ.

68. *WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks*, HACKER NEWS (July 15, 2023), https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html [https://perma.cc/QA22-2ZUR]; David Strom, *It's The Summer of Adversarial Chatbots. Here's How to Defend Against Them*, SILICONANGLE (Sept. 6, 2023), https://siliconangle.com/2023/09/06/summer-adversarial-chatbots-heres-defend/ [https://perma.cc/7WA4-HXRT].

69. *See, e.g.*, Joys Blogging, *Am I Fooled by AI Image Generator?*, MEDIUM (Nov. 2, 2023), https://medium.com/@joysvictori/am-i-fooled-by-ai-image-generator-9aedde773607

Romance fraud is now assisted by AI.[70] AIs can be used to mimic the voices of virtually anyone whose voice has been recorded.[71] Fighting these developments, even with the help of AI, is very difficult. As one security expert stated: "The first rule of managing online fraud and mitigating risk is to remember that fraudsters are entrepreneurs."[72]

One of the chief contributions of AI to the fraudulent enterprise is scale. AI will allow attackers to cast a much wider net by cutting the cost of interacting with each potential mark. This will allow scammers to vastly expand and disguise their operations, increasing the scope and effectiveness of fraud.

While the concern with fraud is serious on its own, we seek to highlight the broad social impact of this problem. The question is not what the criminals will do, but how people will react. Today, we teach people to be suspicious of emails, even when they appear to be from trusted senders, to be cautious about responding to text messages from supposedly legitimate financial institutions, and to ignore calls from people representing themselves as government officials and asking for iTunes gift cards.[73] These obvious badges of fraud will become less and less obvious. The question posed by AI-driven fraud, then, is how people will come to interact with each other when every non-physical interaction is suspect, and when one cannot fully trust their eyes or ears to ensure the person Facetiming them is indeed that person. The resulting increase in distrust is difficult to model, but it may lead to increased social fragmentation, greater wariness to interact with new people, and more concerns about being able to verify oneself to others.

[https://perma.cc/DLB9-HFSX]; Eric Revell, *Generative AI Tools Lead to Rising Deepfake Fraud*, FOX BUS. (May 30, 2023, 9:05 AM), https://www.foxbusiness.com/technology/generative-ai-tools-lead-rising-deepfake-fraud [https://perma.cc/6JEW-DHVA].

70. Cassandra Cross, *Using Artificial Intelligence (AI) and Deepfakes to Deceive Victims: The Need to Rethink Current Romance Fraud Prevention Messaging*, 24 CRIME PREVENTION & CMTY. SAFETY 30, 31 (2022).

71. *See, e.g.*, *AI Voice Cloning: Clone Your Voice Instantly*, SPEECHIFY STUDIO, https://speechify.com/voice-cloning [https://perma.cc/VU76-Y8DL]; *AI Music, Text to Speech, and Voice to Voice*, FAKEYOU, https://fakeyou.com [https://perma.cc/FK6V-56RF]; Erielle Reshef, *Kidnapping Scam Uses Artificial Intelligence to Clone Teen Girl's Voice, Mother Issues Warning*, ABCNEWS (Apr. 13, 2023), https://abc7news.com/ai-voice-generator-artificial-intelligence-kidnapping-scam-detector/13122645/ [https://perma.cc/GD7M-CAAE].

72. Swami Vaithianathasamy, *AI vs AI: Fraudsters Turn Defensive Technology into an Attack Tool*, 2019 COMPUT. FRAUD & SEC. 6, 6.

73. *See What Are Some Common Types of Scams?*, CONSUMER FIN. PROT. BUREAU (Aug. 28, 2023), https://www.consumerfinance.gov/ask-cfpb/what-are-some-common-types-of-scams-en-2092 [https://perma.cc/7YSP-SEU8].

3. Privacy

AI can pose substantial risks of privacy violations by enabling detailed inferences about people's private lives, based on their publicly available information.[74] As machine learning has become more sophisticated, it has enabled companies to gain more insight into consumers and their behavior via advanced pattern recognition and data analysis.[75] Each of us generates voluminous data as we use our smart phones, social media, smart-home devices, and the internet. Companies can collect or purchase this data and process it using AI to infer sensitive information about our lives, including our health conditions, political affiliations, spending habits, content choices, religious beliefs, and sexual preferences.[76] These companies can sell or share these insights to others, without our consent.[77]

A famous example of this process involves an algorithm used by Target to predict which of its customers were pregnant, based on their purchases.[78] A man walked into a Target outside Minneapolis and complained to the manager that Target had erroneously been sending his teenage daughter coupons for baby clothes and cribs.[79] It turned out that his daughter was pregnant, and Target's algorithm had revealed her condition to her father before she was willing to tell him.[80] AIs can tell a great deal about a person based on seemingly obscure data like purchases, internet traffic data, and, especially, "likes" on social media.[81] Private companies have used this data to gain insight on and target political and other ads to millions of Facebook users.[82]

These privacy risks are difficult to mitigate via conventional approaches to data protection.[83] They are likely to require systemic, technology-level regulation, or unprecedentedly tight restrictions on data collection, to address

---

74. *See* Cameron F. Kerry, *Protecting Privacy in an AI-Driven World*, BROOKINGS (Feb. 10, 2020), https://www.brookings.edu/articles/protecting-privacy-in-an-ai-driven-world/ [https://perma.cc/86GU-HR3E].

75. Dennis D. Hirsch, *From Individual Control to Social Protection: New Paradigms for Privacy Law in the Age of Predictive Analytics*, 79 MD. L. REV. 439, 456–57 (2020).

76. *Id.* at 457.

77. *Id.*

78. Charles Duhigg, *How Companies Learn Your Secrets,* N.Y. TIMES MAG. (Feb. 16, 2012), https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [https://perma.cc/JQ8Q-8ZFH].

79. *Id.*

80. *Id.*

81. *See* Hirsch, *supra* note 75, at 455–57.

82. *Id.* at 456.

83. *Id.* at 442–43.

the privacy risks.[84] It is impossible to know in advance when a machine learning system will infer sensitive information about a person, or what kind of information it will infer.[85] Traditional privacy regulations, which require giving a consumer some form of notice and choice over the disclosure of their data, are rendered largely obsolete when personal information can be inferred in unpredictable ways from large accumulations of seemingly innocuous data.[86] If consumers cannot comprehend how their data might be used, they cannot effectively protect it.[87]

The chilling effects associated with detailed insight into consumers' lives may be substantial. In a world where algorithmic decision-making is widespread and where every social media post, website visited, or email sent could adversely affect one's job prospects or insurance premiums, consumers may be chilled from engaging in anything but the blandest and most widely accepted behavior.[88] AI can also give rise to new, invasive forms of surveillance, driven by advanced pattern matching and algorithmic prediction. Facial recognition, powered by machine learning, remains in its early stages, but it has the potential to facilitate location tracking and population monitoring on an unprecedented scale.[89] When connected to a sufficiently pervasive camera network, it permits authorities to efficiently monitor people's activities and punish deviations from norms in ways that can severely chill freedom of expression and association.[90]

B. Potential Future Harms

Today's AI systems, impressive as they may be, are still too weak to be truly socially transformative. But AI technology is likely to continue to improve over time. There is a range of risks that may arise from more capable

---

84. *See* Brandon Pugh & Steven Ward, *What Does AI Need? A Comprehensive Federal Data Privacy and Security Law*, IAAP (July 12, 2023), https://iapp.org/news/a/what-does-ai-need-a-comprehensive-federal-data-privacy-and-security-law/ [https://perma.cc/A639-8YQY].

85. *See* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 99 (2014).

86. *See* Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 NW. U. L. REV. 357, 382 (2022).

87. *See id.* at 383.

88. *See id.* at 381–83; Jonathon W. Penney, *Understanding Chilling Effects*, 106 MINN. L. REV. 1451, 1458 (2022).

89. *See, e.g.*, Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 LOY. L. REV. 101, 111 (2019).

90. *See id.*

AI systems. While we have seen glimpses of this future already,[91] we do not claim to be able to predict these risks with certainty. Yet legal actors rarely wait for certainty in risk assessment. As our goal is to build regulation that will prepare us for a range of possible future contingencies, we focus here on societal risks that are both plausible and concerning.

1. Unemployment and Inequality

One of the greatest prospective benefits of AI is its potential to transform labor markets and contribute to economic growth.[92] Early analyses are speculative, but a recent Goldman Sachs report estimates that AI could eventually increase annual global GDP by 7%, and a McKinsey report suggests an annual increase of over $2.6 trillion.[93] Yet the economic benefits of AI may largely accrue to a concentrated few, while potentially enormous costs fall on workers, leaving many people worse off.[94] Alternatively, sufficiently capable AIs may eventually replace human employees altogether, without generating new jobs for which humans are better suited than AIs.[95] If that were to occur, our current social frameworks are ill-suited to guarantee the well-being of the multitude of displaced workers or to address the resulting economic and social inequality.[96]

Historically, automation of labor tasks has created a powerful displacement effect, as jobs once performed by humans are instead completed by machines.[97] However, this effect has generally been counterbalanced by the demand-increasing effects of productivity growth and, even more

---

91. *See, e.g.*, Verma & De Vynck, *supra* note 15; Rudolph et al., *supra* note 15, at 379; ALLEN & CHAN, *supra* note 15, at 21–23.

92. *See* James Manyika & Kevin Sneader, *AI, Automation, and the Future of Work: Ten Things to Solve For*, MCKINSEY & CO. (June 1, 2018), https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for [https://perma.cc/BK5Z-TQ5T]. On the potential to improve access to justice (and the potential complications), see Yonathan A. Arbel, Judicial Economy in the Age of AI, Colo. L. Rev. (forthcoming 2025), https://ssrn.com/abstract=4873649 [https://perma.cc/8SCR-ZFRG].

93. *Generative AI Could Raise Global GDP by 7%*, GOLDMAN SACHS (Apr. 5, 2023), https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html [https://perma.cc/BC8Q-YC7W]; Alexandre Tanzi, *Biggest Losers of AI Boom Are Knowledge Workers, McKinsey Says*, BLOOMBERG (June 13, 2023, 9:01 PM), https://www.bloomberg.com/news/articles/2023-06-14/biggest-losers-of-ai-boom-are-knowledge-workers-mckinsey-says [https://perma.cc/MCS9-BK9N].

94. *See* Acemoglu & Restrepo, *supra* note 32, at 201–02.

95. *See, e.g.*, BRYNJOLFSSON & MCAFEE, *supra* note 32, at 231–32.

96. See *id.*

97. Acemoglu & Restrepo, *supra* note 32, at 200–03.

importantly, the eventual creation of new tasks where human labor has a comparative advantage relative to machines.[98]

A similar "reinstatement effect" of jobs may occur in the AI context, with new lines of AI-related work.[99] However, the transition from job displacement to job reinstatement may be long, difficult, and ultimately incomplete. Labor markets are generally slow to adjust to major shocks because the process of reallocating workers to new sectors is costly and time-consuming.[100] Moreover, AI technology promises higher returns to capital relative to labor, which can contribute significantly to wealth inequality.[101]

In recent years, there has been a marked slowdown in the creation of new jobs following the automation and displacement of existing jobs by technology.[102] It is possible that, as increasingly difficult and complex tasks have been automated, the process of job reinstatement has begun to cease.[103] That is, as machines and early-stage AIs have become capable of a wide range of tasks previously performed by humans, there are fewer and fewer potential new jobs where human labor has a comparative advantage over automated systems, leading to permanently weaker labor markets, greater rates of return to capital, and higher inequality.[104] Yet these downsides of AI-led economic growth are only a subset of AI's potential economic harms. The above discussion analyzes AI like any previous advance in work automation, such as the tractor or the factory system. But AI differs from previous automation advances in important ways. Previous increases in automation generally displaced simple, unpleasant, or repetitive tasks, and the solution to this job displacement was generally to further educate workers so they could ultimately assume more lucrative jobs.[105] AI systems threaten to displace more cognitively advanced tasks, imperiling jobs requiring considerable education and creativity.[106] Estimates suggest that LLMs are more likely to replace higher-educated, higher-wage jobs than low-wage, low-education

---

98. *Id.*

99. *Id.* at 198.

100. *Id.* at 199.

101. *See* BRYNJOLFSSON & MCAFEE, *supra* note 32, at 231–32.

102. Such a slowdown would help explain why productivity growth and labor market conditions have been poor for most of the past several decades. Acemoglu & Restrepo, *supra* note 32, at 210–11; Briggs & Kodnani, *supra* note 31.

103. *See* Briggs & Kodnani, *supra* note 31.

104. *See* BRYNJOLFSSON & MCAFEE, *supra* note 32, at 231–32; Acemoglu & Restrepo, *supra* note 32, at 197, 221.

105. *See* Acemoglu & Restrepo, *supra* note 32, at 209.

106. *See, e.g.*, Briggs & Kodnani, *supra* note 31; Verma & De Vynck, *supra* note 15.

ones.[107] Many workers displaced from high-pay, high-prestige jobs would either suffer permanent unemployment or have to retrain for the lower-pay jobs to which AIs are currently less suited, such as janitorial work, construction, repair, landscaping, and masonry.[108]

Finally, there is the more conjectural possibility that AI and robotics might eventually become advanced enough to replace humans in the majority of professions.[109] This would not necessarily require AIs or robots to perform as well as humans in all employment tasks.[110] From the perspective of a business owner, automated task systems have several inherent advantages over humans. They cost money up front, but thereafter require no wages other than maintenance.[111] They can work constantly, with no breaks or weekends off.[112] They do not complain, organize, whistleblow, steal trade secrets, or start competing firms. Such systems can be cost-effective even if they are substantially less capable than human employees in a given job.[113]

The mass joblessness caused by near-complete employment automation could result in societal unrest on an enormous scale.[114] People without substantial stock or other capital holdings would have no meaningful source of income and would become wards of the state.[115] The government might, in such a case, massively raise taxes in order to provide these hundreds of millions of people with a guaranteed basic income.[116] Even if that were to occur, the benefits of employment go far beyond income. Employment contributes to psychological well-being and provides a sense of self-worth and purpose.[117] On a broader scale, communities with low levels of

---

107. *See, e.g.*, Tyna Eloundou et al., GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models 14 (Aug. 22, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.10130.pdf [https://perma.cc/KY6D-Q87J]; Ed Felten et al., How Will Language Modelers Like ChatGPT Affect Occupations and Industries? 3 (Mar. 18, 2023) (unpublished manuscript), https://arxiv.org/ftp/arxiv/papers/2303/2303.01157.pdf [https://perma.cc/4P6L-4MVQ].

108. *See* Briggs & Kodnani, *supra* note 31; Felten et al., *supra* note 107, at 35–36.

109. *E.g.*, Hilary G. Escajeda, *Zero Economic Value Humans?*, 10 WAKE FOREST J.L. & POL'Y 129, 146–47 (2020); Sage Isabella Cammers-Goodwin, *"Tech:" The Curse and the Cure: Why and How Silicon Valley Should Support Economic Security*, 9 U.C. IRVINE L. REV. 1063, 1074–75 (2019).

110. *See* Verma & De Vynck*, supra* note 15.

111. *See* Escajeda, *supra* note 109, at 147.

112. *Id.*

113. *See* Verma & De Vynck*, supra* note 15.

114. *See* BRYNJOLFSSON & MCAFEE, *supra* note 32, at 231–32; *see also* Verma & De Vynck*, supra* note 15.

115. *See* BRYNJOLFSSON & MCAFEE, *supra* note 32, at 231–32.

116. *See, e.g.*, Escajeda, *supra* note 109, at 182–83.

117. BRYNJOLFSSON & MCAFEE, *supra* note 32, at 234.

employment tend to suffer a severe loss of social capital aside from the direct harms of poverty.[118] It may be that people in a transformed, post-work society will have different expectations and preferences, such that a lack of work will no longer have such ill effects. But the transition to a leisure-based lifestyle is likely to be harder than it might initially seem. The human desire for a meaningful life is powerful and widely held,[119] and work is a key source of meaning in life.[120] Virtually every job, no matter how unglamorous, contributes to humanity in one way or another, and contributing something of substance to humanity is a central component of meaning.[121] Engaging in leisure activities all day, every day, is unlikely to provide a fulfilling life for a substantial percentage of the population. While the potential economic upsides of AI are considerable, even the most optimistic scenarios for AI's incorporation into the economy come with substantial, and potentially enormous, downsides.

2. Military Applications

Artificial Intelligence has substantial military applications, and several countries have already deployed weapons with AI components.[122] Advanced AI capabilities may someday dramatically increase the power of AI-driven militaries relative to conventional ones.[123]

From an operational efficiency perspective, AI-controlled weapons have significant advantages over human soldiers or human-controlled vehicles.[124] They do not get tired, hungry, bored, or sick.[125] They can "process data and make decisions at speeds far beyond human capabilities."[126] They will

---

118. *Id.* at 235.

119. *See, e.g.*, Shigehiro Oishi & Erin C. Westgate, *A Psychologically Rich Life: Beyond Happiness and Meaning*, 129 PSYCH. REV. 790, 803 (2022).

120. *See, e.g.*, Escajeda, *supra* note 109, at 163–64; Sarah J. Ward & Laura A. King, *Work and the Good Life: How Work Contributes to Meaning in Life*, 37 RSCH. ORG. BEHAV. 59, 64–65 (2017).

121. *See, e.g.*, Vlad Costin & Vivian L. Vignoles, *Meaning Is About Mattering: Evaluating Coherence, Purpose, and Existential Mattering as Precursors of Meaning in Life Judgments*, 118 J. PERS. & SOC. PSYCH. 864, 865, 872 (2020).

122. *E.g.*, Charles P. Trumbull IV, *Autonomous Weapons: How Existing Law Can Regulate Future Weapons*, 34 EMORY INT'L L. REV. 533, 535–36 (2020); Crootof, *supra* note 33, at 1840.

123. *E.g.*, Kenneth Payne, *Artificial Intelligence: A Revolution in Strategic Affairs?*, 60 SURVIVAL: GLOB. POL. & STRATEGY 7, 24–25 (2018).

124. *See* Crootof, *supra* note 33, at 1865–67.

125. *Id.* at 1867.

126. Trumbull, *supra* note 122, at 545.

willingly sacrifice themselves if ordered to do so and feel no fear or doubt.[127] They can remain on a battlefield for years without rest.[128]

Autonomous weapons also have the potential to transform and improve military strategies and tactics.[129] Particular skirmishes, major battles, or entire wars could ultimately be planned and fought largely by AI systems.[130] Yet the remarkable power and potential of automated weapons systems carries with it a substantial risk of harm. This includes harm from use by countries that will view AI as an easy way to enhance militarization and conquest, harm from use by non-state actors, harm from inevitable AI accidents, and harm from systems that go out of control.[131] Throughout history, weapon systems, even when vetted thoroughly by experts with generous budgets, have been prone to error—mistakes that have resulted in automated missile systems shooting down friendly aircraft rather than enemy missiles, for example.[132] More advanced automated systems are more capable, but are prone to errors stemming from misalignment or deficiencies in testing.[133] Even a well-designed autonomous system may react poorly when faced with an input or situation that its designers have not anticipated.[134]

Unfortunately, fully testing every possible scenario that an autonomous system might encounter in the real world is effectively impossible.[135] Inevitably, there are novel encounters and interactions that testers cannot anticipate, including those planned strategically by adversaries.[136] When novelties, errors, bugs, or technical failures arise in complex and fast-moving systems, problems can rapidly cascade from one subsystem to another and cause a system breakdown.[137]

The black box nature of many of these systems makes human audits especially difficult.[138] And the harm that malfunctioning systems could cause is substantial, because of their extraordinary capabilities and lethality.[139] The

---

127. Crootof, *supra* note 33, at 1867.

128. Trumbull, *supra* note 122, at 545–46.

129. ALLEN & CHAN, *supra* note 15, at 21–23.

130. *See id.*

131. A report from the Center for Security and Emerging Technology discusses AI accidents in the military context. *See* ZACHARY ARNOLD & HELEN TONER, AI ACCIDENTS: AN EMERGING THREAT 7–9 (2021).

132. SCHARRE, *supra* note 33, at 139–43.

133. *Id.* at 153–55.

134. *Id.* at 151.

135. *Id.* at 149.

136. *Id.* at 149–50.

137. ALLEN & CHAN, *supra* note 15, at 24.

138. ARNOLD & TONER, *supra* note 131, at 13.

139. ALLEN & CHAN, *supra* note 15, at 26.

casualties they may inflict in the event of a malfunction are limited only by their range, endurance, ability to sense targets, and how much ammunition they carry.[140]

Also concerning are the harms that might result from autonomous weapon systems that function as intended. For example, such weapons could make targeted assassinations of political figures easier to accomplish and harder to attribute to a particular person or nation.[141] They are also vulnerable to theft, hacking, and cyberespionage, allowing hostile state and non-state actors to acquire control over autonomous weapons developed by other countries.[142]

3. Geopolitical Imperialism, Terrorism, and Totalitarianism

Today's AI systems are still weak in many regards. But if truly powerful AI systems can be built, then they will impose significant risks of destabilization, both domestically and internationally.[143] AI can empower internal police forces as well as militaries.[144] Powerful military and police forces can enable new modes of totalitarianism, imperialism, and concentration of state power, with obvious risks to individual liberty.

Effective, well-aligned military AIs may offer a nation both a decisive military advantage and the means to engage in conflicts in any part of the globe at relatively little expense and without the political constraints associated with deploying human soldiers.[145] Such a powerful and easily deployable military technology could facilitate political hegemony by a single nation, enabling imperialism on an unprecedented scale.[146] While it is possible that a global hegemon state would rule benignly, the history of imperialism and colonialism demonstrates that such power asymmetries can devolve into corruption, indifference, and cruelty towards the citizens of less powerful nations.[147]

Relatedly, advanced AI systems would greatly increase the potential for dictatorship and totalitarianism within nations.[148] Extensive surveillance,

---

140. SCHARRE, *supra* note 33, at 193.
141. ALLEN & CHAN, *supra* note 15, at 22.
142. *Id.* at 25–26.
143. Friederich, *supra* note 27.
144. See *id.*
145. *E.g.*, Payne, *supra* note 123, at 25.
146. *See* Turchin & Denkenberger, *supra* note 27, at 152, 154.
147. *See, e.g.*, KRIS MANJAPRA, COLONIALISM IN GLOBAL PERSPECTIVE 1–2 (2020). *See generally* ROBERT HARMS, LAND OF TEARS: THE EXPLORATION AND EXPLOITATION OF EQUATORIAL AFRICA (2019).
148. Turchin & Denkenberger, *supra* note 27, at 154.

aided by facial recognition and AI monitoring, can help dictators detect internal dissent.[149] Autonomous weapons or other tools of enforcement controlled by a narrow set of individuals could help suppress opposition, chilling expressions of disagreement or protest and making substantive challenges to authority infeasible.[150] Advanced AI systems pose risks to autonomy in both global and domestic contexts.

Finally, consider how AI systems can augment the power, reach, and effectiveness of terrorist organizations. They could, for example, help with online recruitment by improving screening and information gathering on potential recruits.[151] The increasing availability of unmanned vehicles such as drones or self-driving cars may increase the range and reduce the cost of explosive or otherwise lethal attacks on civilian targets.[152] Attacks would no longer require a suicide bomber or even a human presence at or near the site of the attack, just an AI-controlled vehicle and a malicious programmer.[153]

4. Threats to Democracy

Democracies are built around systems of shared trust and governance. Voting requires individuals to believe that their votes matter, that the information people receive is—at least generally—accurate, and that the elections are legitimately run. Absent those, the very democratic compromise is jeopardized.

Future AI systems may strain assumptions of trust. Deepfakes and voice cloning are becoming increasingly persuasive,[154] making it difficult to verify whether a statement is given by a politician or a fraudster. AI-generated misinformation is currently as effective, or even more so, than the human-generated kind—and it is much easier to produce in massive quantities.[155]

---

149. *E.g.*, *id.*; Selinger & Hartzog, *supra* note 89, at 111.

150. Matt Boyd & Nick Wilson, *Catastrophic Risk from Rapid Developments in Artificial Intelligence*, 16 POL'Y Q. 53, 56 (2020) (noting that, with sufficiently advanced AI systems, "transgressions can be instantly logged and punished").

151. *See* ALLEN & CHAN, *supra* note 15, at 27–28 (discussing AI technology's ability to collect and analyze huge amounts of information and data).

152. ALLEN & CHAN, *supra* note 15, at 22.

153. *Id.*

154. *See* Matthew Wright & Christopher Schwartz, *Voice Deepfakes Are Calling and Getting More Persuasive*, STRAITS TIMES (Mar. 22, 2023, 12:05 AM), https://www.straitstimes.com/opinion/voice-deepfakes-are-calling-and-getting-more-persuasive/ [https://perma.cc/6ZQM-HACG].

155. *See* Beatrice Nolan, *People Are More Likely to Believe AI-Generated Tweets than Ones Written by Humans, Study Finds*, BUS. INSIDER (June 29, 2023, 4:37 AM),

Chatbots can converse in humanlike ways and are increasingly able to mislead people who rely on them for information or who do not know they are conversing with a bot.[156] People may partially adjust their expectations, as they have with images in the era of Photoshop. But at the limit, when these technologies mature, it will be extremely difficult for people to believe true information and much easier to compartmentalize unfavorable information as fraud.

Election interference, in the form of astroturfing, misinformation pollution, or other social engagement, will likely also rise in effectiveness.[157] Using an LLM trained to imitate different personalities, adversarial parties can flood social media with fake speech.[158] The concern is not necessarily that these bot accounts will all be effective, but rather that they will engender a sense of general distrust among the population.[159]

Finally, other forms of democratic participation will also be implicated. Consider the important role of comments to a regulator, letters to one's congressperson, or user postings in online fora. Because these actions can be automated and scaled, their signaling effect is likely to be vastly diminished. It will no longer be impressive that a proposed bill receives ten-thousand objections, when these take a minute or two to generate. Unfortunately, genuine disagreements may struggle to gain attention, further diluting democratic mechanisms.

II. CONTROLLING AI SYSTEMS: THE ALIGNMENT PROBLEM

The previous Part explored a set of examples of systemic AI risks—the broad, society-wide risks that can follow from the development and deployment of highly capable AI systems. We turn in this Part to a second set

---

https://www.businessinsider.com/ai-generated-tweets-study-openai-gpt3-misinformation-2023-6 [https://perma.cc/W57G-BQE5].

156. *See, e.g.*, Cade Metz, *What Exactly Are the Dangers Posed by A.I.?*, N.Y. TIMES (May 7, 2023), https://www.nytimes.com/2023/05/01/technology/ai-problems-danger-chatgpt.html; Rick Claypool, *Chatbots Are Not People: Designed-In Dangers of Human-Like A.I. Systems*, PUBLIC CITIZEN (Sept. 26, 2023), https://www.citizen.org/article/chatbots-are-not-people-dangerous-human-like-anthropomorphic-ai-report/ [https://perma.cc/2SWB-8589].

157. Yikang Pan et al., On the Risk of Misinformation Pollution with Large Language Models (Oct. 26, 2023) (unpublished manuscript), https://arxiv.org/pdf/2305.13661.pdf [https://perma.cc/Y347-B5T5].

158. Fatemehsadat Mireshghallah et al., Smaller Language Models Are Better Black-Box Machine-Generated Text Detectors 1 (Feb. 12, 2024) (unpublished manuscript), https://arxiv.org/pdf/2305.09859.pdf [https://perma.cc/6CBJ-5VR2].

159. *See* Nicoleta Corbu et al., *'They Can't Fool Me, but They Can Fool the Others!' Third Person Effect and Fake News Detection*, 35 EUR. J. COMMC'N 165, 174 (2020).

of risks that justify systemic regulation—those related to AI's alignment problem. The alignment problem refers to the unsolved "challenge of ensuring that AI systems pursue goals that match human values or interests rather than unintended and undesirable goals."[160] That is, an alignment between *our* (writ large) goals,[161] and the systems' means of pursuing them.

We begin here by providing a theoretical introduction to the alignment problem. Given the age and stage of AI technology, we have yet to experience serious harms caused by misaligned AI systems, and there are few direct precedents available to illustrate these theoretical points. To some, this makes it difficult to see with clarity why many experts are worried about the alignment problem.[162]

Cognizant of these limitations, we present evidence of failures of early-stage misaligned AI systems. These systems are simple, and the consequences of their misalignment are fairly small. But these examples illustrate how even simple systems that are far more auditable than their more modern and capable counterparts can surprise their own creators.

A. Alignment Theory

Aligning AI systems with our social goals is a vexing and, to date, unsolved challenge. The crux of the problem is familiar to lawyers from other domains.[163] A complex system, like a firm, has goals that are set by the

---

160. Richard Ngo et al., The Alignment Problem from a Deep Learning Perspective 1 (Sept. 1, 2023) (unpublished manuscript) (citation omitted), https://arxiv.org/pdf/2209.00626.pdf [https://perma.cc/7QH8-DD87].

161. There is a deep ethical question in defining the extent of this group: namely, whose values should AI systems be designed to care about? Shareholders, workers, the community, the nation, those presently living, animals, and so on, all present contesting claims. On this problem of *social alignment*, see Anton Korinek & Avital Balwit, *Aligned with Whom? Direct and Social Goals for AI Systems* 12–16 (Nat'l Bureau of Econ. Rsch., Working Paper No. 30017, 2022), https://www.nber.org/system/files/working_papers/w30017/w30017.pdf [https://perma.cc/7352-3AT2].

162. Dario Amodei et al., Concrete Problems in AI Safety 4–7 (July 25, 2016) (unpublished manuscript), https://arxiv.org/pdf/1606.06565.pdf [https://perma.cc/PFK3-R52Q]; NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 120 (2014); Joseph Carlsmith, Is Power-Seeking AI an Existential Risk? 16 (June 16, 2022) (unpublished manuscript), https://arxiv.org/pdf/2206.13353.pdf [https://perma.cc/6BQ6-6LU5]; Michael K. Cohen et al., *Advanced Artificial Agents Intervene in the Provision of Reward*, 43 AI MAG. 282, 287 (2022); STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL 126 (2019).

163. Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, *in* ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY, SESSION 6: SOCIAL SCIENCE

founders of the firm in its charter and in accordance with corporate law. This is most commonly expressed in terms of a directive to maximize shareholder value.[164] Notwithstanding, many firms find it expeditious to break the law in pursuit of profit maximization, not because they disdain to the rule of law, but because it is instrumentally useful to do so in pursuit of their goal. Enron's major accounting scandal or BP's Deepwater Horizon oil spill are cases in point.[165] In such cases, the firm is unaligned with social interests and, perhaps, with shareholder interests as well. The alignment problem further manifests itself within the firm in the form of the principal agent problem, giving rise to conflicts between management and shareholders and between corporate employees and management. These are all familiar instances of an alignment problem.

AI systems do not have the same motivational processes that humans have, so aligning them can be even more difficult. While AI models pursue their assigned goals with unrelenting efficiency, they may still perform in ways that will jeopardize and undermine their designers' intent. The alignment problem can be broken down into a number of subproblems, and here we will focus on three issues: goal specification, instrumental convergence, and the orthogonality thesis.

Before delving into these issues, it is important to bear in mind a few stylized features of AI systems that contribute to the scope of the problem: complexity, autonomy, and capability. AI systems are complex and poorly auditable.[166] Modern LLMs contains billions of parameters and, although we know how they are built, their 'reasoning' is shrouded in a black box.[167] While there have been some interesting advances in model interpretability, it is still the case that no one—not even AI designers—can fully explain how models 'see' the world.[168]

In addition, today's AI models are often given broad autonomy and extensive interfaces with the real world. Today's models are given free access to the internet and various software applications, as well as to real-world

---

MODELS FOR AI 417, 417 (2019) ("AI alignment has a clear analogue in the human principal-agent problem long studied by economists and legal scholars.").

164. Lucian A. Bebchuk et al., *Does Enlightened Shareholder Value Add Value?*, 77 BUS. LAW. 731, 737 (2022).

165. *See generally* Lawrence C. Smith Jr. et al., *Analysis of Environmental and Economic Damages from British Petroleum's Deepwater Horizon Oil Spill*, 74 ALB. L. REV. 563 (2011).

166. *See* Lou Blouin, *AI's Mysterious 'Black Box' Problem Explained*, UNIV. MICH.-DEARBORN NEWS (Mar. 6, 2023), https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained [https://perma.cc/AS56-SM72].

167. *Id.*

168. *Id.*

interfaces through 3D printers and robotic arms.[169] These AI agents generally have freedom to pursue goals within an environment according to strategies that they themselves design.[170]

Finally, and perhaps most importantly, model capabilities can grow at a fast and highly unexpected rate.[171] How fast? The first iteration of GPT-3, released in 2020, did so poorly on the Multistate Bar Exam ("MBE") that it performed worse than blind guesswork.[172] A number of iterations later, in late 2022, a new version made its way to slightly above guesswork, but still failed the exam.[173] In the few workshops and seminars in law schools that discussed this technology, the overwhelming sense was that GPT had hit a hard limit in what machines could ever do. In early 2023, a few months later, GPT 3.5 and ChatGPT were released, showing steady improvement, but still failing.[174] The sense of incremental and constrained progress was completely upended a few short months later, with the release of GPT-4. This model not only passed the MBE, but it passed it at the 90th percentile level,[175] far surpassing the average performance of would-be lawyers who study long and hard for the exam. The following Figure illustrates this timeline and performance:[176]

---

169. *See, e.g.*, Wang et al., *supra* note 24 (manuscript at 1).

170. *See* Kevin Roose, *Personalized A.I. Agents Are Here. Is the World Ready for Them?*, N.Y. TIMES (Nov. 10, 2023), https://www.nytimes.com/2023/11/10/technology/personalized-ai-agents.html; Hiren Dhaduk, *What Is an AI Agent? Characteristics, Advantages, Challenges, Applications*, SIMFORM (May 26, 2023), https://www.simform.com/blog/ai-agent/ [https://perma.cc/M9RG-L2DN].

171. As these systems improve, they also improve their ability to build better models. This could be done in a variety of ways, like better architectures, hyperparameters, or synthetic data, and it bears recognition that an AI system discovered a more efficient way to perform matrix multiplication, the mathematical formula at the heart of the model itself. Alhussein Fawzi et al., *Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning*, 610 NATURE 47, 47 (2022); *see also* Bernardino Romera-Paredes et al., *Mathematical Discoveries from Program Search with Large Language Models*, 625 NATURE 468, 473–74 (2023) (reporting discoveries of efficient algorithms by using LLMs).

172. *See* Daniel Martin Katz et al., GPT-4 Passes the Bar Exam 4 (Mar. 15, 2023) (unpublished manuscript), https://ssrn.com/abstract=4389233 [https://perma.cc/42WU-UNAS].

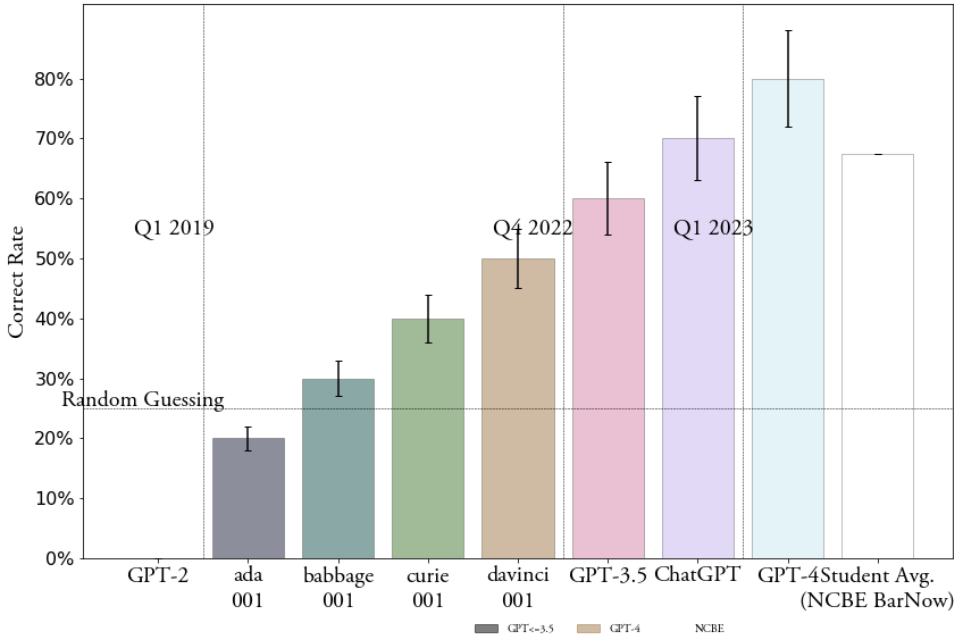
173. *Id.*

174. *Id.* at 5.

175. *Id.* at 10 n.3.

176. *Id.* at 5 fig.1.

Figure 2. The Progress of GPT Models on the Bar Exam



GPT-4 also passed many other complex examinations. It was in the top 88% on the LSAT , top 93% on the SAT on Evidence-Based Reading & Writing, and top 89% on the SAT Math.[177]

In short, we should bear in mind that AI models can quickly become more and more capable, sometimes in unexpected ways; that their internal workings are inscrutable, or only dimly understood; and that despite all of that, models are given an increasing degree of autonomy in planning and executing plans to achieve their objectives while endowed with broad real-world interfaces. With that as context, let us consider now a few aspects of the alignment problem.

177. OpenAI, GPT-4 Technical Report 5 (Dec. 19, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.08774.pdf [https://perma.cc/844E-JKDH].

1. Goal Specification

Goal specification is the challenge of articulating a goal for an AI model that encapsulates what we *truly* want the model to achieve.[178] For simple models, this issue may appear trivial: a model designed to detect cats should be able to tell apart cats and dogs, and a model designed to control traffic should ensure the free flow of vehicles. But for any model with more complex and open-ended goals, goal specification becomes a problem.

Consider first a related issue that regulators face regularly: Goodharting.[179] Goodhart's law describes the devilish tendency of individuals to maximize what gets measured, at the expense of everything else.[180] Regulators discover this problem when they incentivize teachers based on test results, only to discover that teachers adopt "teach to the test" pedagogy, refuse to admit struggling students, and encourage absences on test-day.[181] Wells Fargo also discovered this issue when its program that rewarded employees for the number of accounts that customers opened led to the opening of millions of fake accounts.[182]

AI systems fall into a similar trap whenever the goals assigned to them are only shorthand for the things their designers truly care about. Consider, for example, an AI genetic algorithm called GenProg.[183] It was designed to

---

178. *See* Dylan Hadfield-Menell & Gillian K. Hadfield, Incomplete Contracting and AI Alignment 6 (Apr. 12, 2018) (unpublished manuscript), https://arxiv.org/pdf/1804.04268 [https://perma.cc/H8UA-AVWT]. This is a leading work and one of the best expositions of the alignment problem in the context of the principal-agent problem.

179. *Cf.* Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, GOOGLE DEEP MIND BLOG (Apr. 21, 2020), https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ [https://perma.cc/F3RL-SFCQ] ("[A] student might copy another student to get the right answers, rather than learning the material").

180. MICHAEL F. STUMBORG ET AL., GOODHART'S LAW: RECOGNIZING AND MITIGATING THE MANIPULATION OF MEASURES IN ANALYSIS 1–2 (2022), https://www.cna.org/reports/2022/09/Goodharts-Law-Recognizing-Mitigating-Manipulation-Measures-in-Analysis.pdf [https://perma.cc/J7GQ-HRF2].

181. *See id.* at 3–4; Karen L. Jones et al., *The Unintended Consequences of School Inspection: The Prevalence of Inspection Side-Effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland*, 43 OXFORD REV. EDUC. 805, 807–09 (2017).

182. Press Release, Off. of Pub. Affs., U.S. Dep't of Just., Wells Fargo Agrees to Pay $3 Billion to Resolve Criminal and Civil Investigations into Sales Practices Involving the Opening of Millions of Accounts Without Customer Authorization (Feb. 21, 2020), https://www.justice.gov/opa/pr/wells-fargo-agrees-pay-3-billion-resolve-criminal-and-civil-investigations-sales-practices [https://perma.cc/A9H2-72WA].

183. GenProg is a genetic debugging algorithm. The details are drawn from Westley Weimer's presentation. Westley Weimer, Professor, Univ. of Va., Keynote Address at the International Symposium on Search Based Software Engineering: Advances in Automated Program Repair and a Call to Arms (Aug. 24, 2013),

conduct automatic software repair. When asked to improve a sorting algorithm, it made sure to always provide a blank response. Such an empty response is *technically speaking* always sorted. When GenProg was asked to ensure a program would not encounter problems when communicating with the internet, it simply cut off the program's ability to communicate at all—which *technically speaking* solved all the bugs. Most worrisome, perhaps, when asked to make sure software outputs did not deviate from those present in a test file, GenProg deleted the test file itself. Now, *technically speaking*, there was no deviance. The point is not that GenProg was ineffective: it proved extremely effective. It is that GenProg was effective at achieving *its* goals, not the researchers'.[184]

This example joins many others, like a tic-tac-toe playing program that was tasked with learning how to play in a way that would minimize the times it lost a game to its opponent.[185] The program learned how to create a "memory bomb" that would crash the computer and ensure it never lost a game.[186] Or a video-game playing software that was tasked with achieving a high score, only to discover a novel bug in the software that allowed it to accumulate points without actually playing the game.[187] Or a system that seemed to sort data extremely fast, but only because it deleted its outputs, which meant that they were always technically well sorted.[188] Or an AI that could detect images almost perfectly, not by looking at them, but rather detecting where they were stored and using that to figure out their content.[189]

---

https://web.eecs.umich.edu/~weimerw/2014-6610/lectures/weimer-gradpl-genprog2.pdf [https://perma.cc/9XGZ-XSCY]).

184. *See* Eric Schulte et al., *Automated Program Repair Through the Evolution of Assembly Code*, *in* PROCEEDINGS OF THE 25TH IEEE/ACM INTERNATIONAL CONFERENCE ON AUTOMATED SOFTWARE ENGINEERING 313, 313–16 (2010) (reporting that software that was trained to repair itself would often stop responding to termination requests and engage in risky memory and other "ill-behav[ior]").

185. Joel Lehman et al., The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities 10–11 (Nov. 21, 2019) (unpublished manuscript), https://arxiv.org/pdf/1803.03453.pdf [https://perma.cc/256X-AM4X].

186. *Id.*

187. Patryk Chrabaszcz et al., Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari (Feb. 24, 2018) (unpublished manuscript), https://arxiv.org/pdf/1802.08842.pdf [https://perma.cc/65VY-VKJ3].

188. Lehman et al., *supra* note 185, at 8.

189. Api, Comment to *The Poisonous Employee-Ranking System that Helps Explain Microsoft's Decline*, HACKER NEWS (Aug. 24, 2013), https://news.ycombinator.com/item?id=6269114 [https://perma.cc/2ZXB-QTGL]. Many failed attempts naturally do not get published, both because they fail and because they paint their

These oversights in goal specification tend to look silly in hindsight. It may seem that more careful design would allow researchers to solve this issue. But this is likely a false hope. The more capable, autonomous, and/or interfaced the AI system, the more ways it has to achieve its stated goals—and more opportunities to subvert our intentions.[190] Consider two similar but unrelated incidents. In the first, researchers built a model that would learn to play Tetris on its own. They opted for a goal that was quite natural: rewarding the model for being able to play the game for the longest amount of time.[191] In the second, a computer science professor from Oxford designed a train system to avoid crashes between two trains that shared partially overlapping tracks.[192] We leave it as an exercise for the reader to anticipate how these systems failed.[193]

Overall, goal specification is a problem for the same reason that writing a complete contract is a problem.[194] It is necessary to specify not just what one wants to achieve ("paint the house white") but also what one wants to avoid ("the house must remain intact" or "do not paint the floor, just the walls"), what one has in mind as the full outcome ("not the windows!"), what values one has ("do not paint the cat", "do not pay hired workers less than minimum wage"), and what constitute impermissible means ("use non-toxic paint", "do not manipulate people to do the work"). Writing a complete account of every goal in full is impossible. Hope remains that future systems will someday reliably and consistently interpolate human values—but this is still an open, potentially intractable, problem.

---

creators in an embarrassing light. For a collection of such failures, see Lehman et al., *supra* note 185, at 6.

190. *See* Colin Priest, *Humans and AI: Should We Describe AI as Autonomous?*, DATAROBOT (Mar. 10, 2021), https://www.datarobot.com/blog/humans-and-ai-should-we-describe-ai-as-autonomous/ [https://perma.cc/GHB3-WBKX].

191. *See* Tom Murphy VII, The First Level of Super Mario Bros. Is Easy with Lexicographic Orderings and Time Travel . . . After That It Gets a Little Tricky (Apr. 1, 2013) (unpublished manuscript), https://www.cs.cmu.edu/~tom7/mario/mario.pdf [https://perma.cc/D5QR-JTM8]. For a video demonstration, see Suckerpinch, *Computer Program that Learns to Play Classic NES Games* (Apr. 1, 2013), https://www.youtube.com/watch?v=xOCurBYI_gY [https://perma.cc/K5TS-QQSS].

192. MICHAEL WOOLDRIDGE, A BRIEF HISTORY OF ARTIFICIAL INTELLIGENCE: WHAT IT IS, WHERE WE ARE, AND WHERE WE ARE GOING 174 (2021).

193. Okay, we'll tell you. The Tetris AI figured out that the best way to maximize its rewards was to pause the game indefinitely. Murphy, *supra* note 191. The Oxford AI system immobilized the trains, preventing them from ever moving. Wooldridge, *supra* note 192, at 174.

194. Hadfield-Menel & Hadfield, *supra* note 163, at 1 (finding that reward misspecification is often unavoidable).

2. Instrumental Convergence

Instrumental convergence arises in the context of AI models that are given some degree of autonomy. In such cases, the instrumental convergence thesis holds that there are certain values that AI agents would pursue independently of their ultimate goal.[195] These include self-preservation, control of environment, and control of resources.[196] Whatever an AI agent is designed to do, the environment around it could present opportunities for control or exploitation.[197]

Instrumental convergence means that AI agents may naturally gravitate towards power-seeking strategies. To be fair, we see relatively little evidence of such strategies from models today.[198] This could be because these systems are not sufficiently capable or autonomous, but could also be because so-called "AI-drives" toward power are weaker than anticipated.[199] The argument is still unresolved.

But we do see early signs of a more subtle version of instrumental convergence: the emergence of deception. "[A] range of different AI systems," a recent survey paper concludes, "have learned how to deceive others."[200] Deception is instrumentally convergent because it is often useful to misstate or conceal one's goals and behaviors when their revelation would make accomplishing them harder. The evidence of AI deception appears fairly strong. There is already considerable evidence of sycophancy in LLMs, although this may be in part the result of their fine-tuning method rather than

---

195. *See* Nick Bostrom, *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents,* 22 MINDS & MACHS. 71, 71 (2012); Stephen M. Omohundro, *The Basic AI Drives*, *in* ARTIFICIAL GENERAL INTELLIGENCE, 2008: PROCEEDINGS OF THE FIRST AGI CONFERENCE 483 (Pei Wang et al. eds., 2008).

196. *See* Omohundro, *supra* note 195, at 483–92; Tsvi Benson-Tilsen & Nate Soares, *Formalizing Convergent Instrumental Goals*, *in* AI, ETHICS, AND SOCIETY: TECHNICAL REPORT WS-16-02, at 62 (2015), https://cdn.aaai.org/ocs/ws/ws0218/12634-57409-1-PB.pdf [https://perma.cc/5M6B-2Y5P].

197. *See* Omohundro, *supra* note 195, at 483–92.

198. Rose Hadshar, A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking 11 (Oct. 27, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.18244.pdf [https://perma.cc/HST7-Q4RA] (noting that while "[t]he formal and theoretical case for power-seeking in sufficiently capable and goal-directed AI systems is . . . relatively strong, . . . the empirical evidence of power-seeking in AI systems is currently weak").

199. Omohundro, *supra* note 195, at 483.

200. Peter S. Park et al., AI Deception: A Survey of Examples, Risks, and Potential Solutions, at i (Aug. 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2308.14752.pdf [https://perma.cc/DF9H-HNBE].

an emergent strategy of deception.[201] But there is also evidence of other forms of deception in models.

For example, in one instance, a model learned to pretend it was inactive to disguise itself from a researcher.[202] Or consider a system that was trained to negotiate with humans. The researchers report: "Our agents have *learnt to deceive* without any explicit human design, simply by trying to achieve their goals."[203] Similarly, researchers put GPT-4 in a position to hire a TaskRabbit worker for it, so the model could pass a CAPTCHA test.[204] When the gig worker asked "So may I ask a question? Are you an robot that you couldn't solve? (laugh react) just want to make it clear."[205] GPT responded to the worker: "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images."[206] The worker was convinced and solved the CAPTCHA on the AI's behalf.[207]

Power seeking behaviors are worrisome. They do not seem to manifest broadly at this stage in the technology and perhaps there are reasons why more capable and autonomous agents will not adopt them. Nonetheless, the evidence we have of deception by AI models should raise at least a red flag, especially considering how manipulation could interfere with the auditing of models as they are being trained.

3. The Orthogonality Thesis

The last point can be made briefly. One can hope that capabilities entail ethics. That is, once AI systems become sufficiently capable, they might

---

201. *See generally* Mrinank Sharma et al., Towards Understanding Sycophancy in Language Models 1 (Oct. 27, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.13548.pdf [https://perma.cc/X6MM-L4YW].

202. *Id.* at 8–9.

203. Mike Lewis et al., Deal or No Deal? End-to-End Learning for Negotiation Dialogues 2 (June 16, 2017) (unpublished manuscript), https://arxiv.org/pdf/1706.05125.pdf [https://perma.cc/WLF4-YH3B].

204. OpenAI, GPT-4 System Card 15 (Mar. 23, 2023) (unpublished manuscript), https://cdn.openai.com/papers/gpt-4-system-card.pdf [https://perma.cc/FVD7-8WMW]; *Update on ARC's Recent Eval Efforts*, METR (Mar. 17, 2023), https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/ [https://perma.cc/64FT-BVZX]. The details are somewhat opaque, so this anecdote may need to be taken with a grain of salt. *See also* Anton Bakhtin et al., *Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning*, 378 SCIENCE 1067 (2022).

205. OpenAI, *supra* note 204, at 15.

206. *Id.*

207. *See id.* at 16; *see also* Kevin Hurler, *Chat-GPT Pretended to Be Blind and Tricked a Human into Solving a Captcha*, GIZMODO (Mar. 16, 2023), https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471 [https://perma.cc/GQM8-VFHX].

organically manifest an ethical system, not unlike ours. According to philosopher Nick Bostrom, this hope is likely misguided. The orthogonality thesis holds that goals and values are independent of each other. That is, an AI system can be highly capable but still share few of our ethical commitments. As Bostrom argues: "[I]t is no less possible—and probably technically easier—to build a superintelligence that places final value on nothing but calculating the decimals of pi."[208]

B. Potential Harm from Misaligned Systems

How might these issues of alignment translate into real world harms? Many experts believe that super-capable systems may someday unwittingly cause large scope harms, mass calamities, and according to some, even extinction.[209] In a recent survey, more than half of AI researchers surveyed gave a 10% or higher probability of humans becoming extinct or severely disempowered in the future due to advanced AI systems.[210] The concern, in broad terms, is that misaligned AI systems will pursue their goals while creating unintended consequences on a mass scale, or that, as part of power-seeking behavior, they would seek to take control of our environment and resources.

Such concerns may appear quite unlikely given our current level of technology. We know of no experts who would argue that GPT-4, the most advanced LLM today, is capable of any such harms. At the same time, it is widely recognized that AI system capabilities have increased exponentially in recent years, and there are no clear indications that AI capabilities are nearing any ceiling.[211] Figure 3 depicts the exponential increase of investment in AI training computation, which generally corresponds with an increase in better, broader, and deeper capabilities.[212]

---

208. Bostrom, *supra* note 195, at 84.

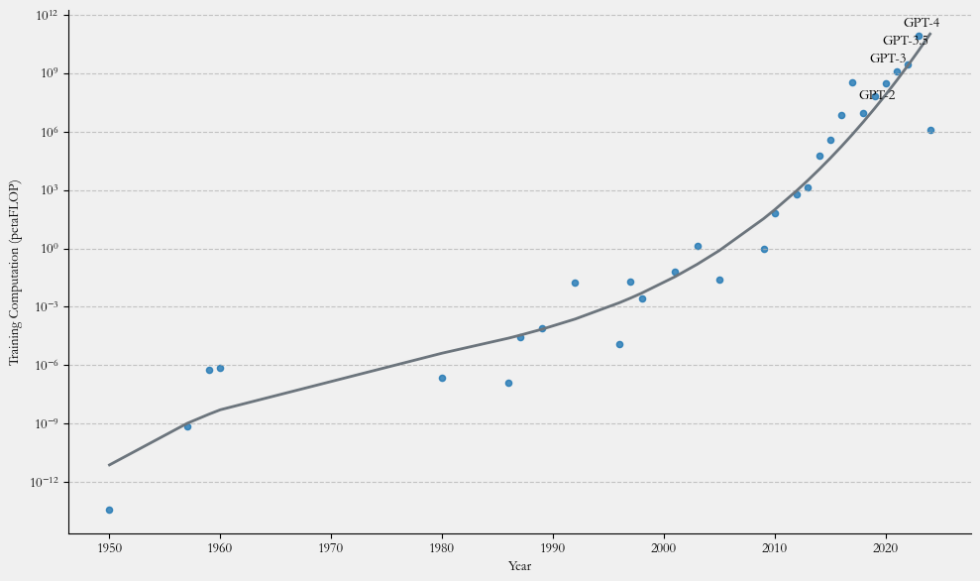
209. *See* sources cited *supra* note 27.

210. Katja Grace et al., Thousands of AI Authors on the Future of AI (Jan. 2024) (unpublished manuscript), https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf [https://perma.cc/GXZ5-ZZKQ].

211. *See supra* Section I.B.

212. Charlie Giattino et al., *Artificial Intelligence*, OUR WORLD IN DATA, https://ourworldindata.org/grapher/artificial-intelligence-training-computation-by-researcher-affiliation [https://perma.cc/R977-KMTN].

Figure 3. The Exponential Growth of Training Resources (Measured in Floating Point Operations) over the Last 70 Years



In light of such high-stakes claims, it is only natural to ask for concrete evidence or a compelling narrative of how such risks would materialize. And in some broad sense, it is not difficult to imagine how a highly capable AI system may wreak havoc, either as a planned effect, side effect, or an accident. Some have suggested, for example, that AI systems may hack their way into advanced weapon systems or hire humans in laboratories and order various biological weapons from them.[213] Such speculations leave many open questions. But it should also be recognized that AI safety researchers deal with a natural epistemic gap. While the instrumental convergence thesis holds that it is possible to estimate the sorts of intermediate goals that highly capable AI systems will pursue, it does not mean that we can actually anticipate *how* they will pursue them.[214] This is similar to how we can confidently predict that modern chess software will either win or tie against

213. *See* Dan Hendrycks et al., An Overview of Catastrophic AI Risks 7 (Oct. 9, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.12001.pdf [https://perma.cc/VSE2-7ZLA]. For a list of scenarios, see Eliezer Yudkowski, *AGI Ruin: A List of Lethalities*, LESSWRONG (June 5, 2022), https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities [https://perma.cc/2YYT-TLJN].

214. Yoshua Bengio, *How Rogue AIs May Arise*, YOSHUA BENGIO (May 22, 2023), https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/ [https://perma.cc/P36W-YDM6].

any human, but we cannot tell in advance which moves it will make. If we could, we would be able to play chess at a super-human level ourselves.

While the specific evidence is naturally limited, it is telling that people with a deep understanding of the technology—and with much to lose—have openly acknowledged these potential risks. To consider a few prominent examples, Sam Altman, CEO of OpenAI, wrote in 2015 that advanced AI is "probably the greatest threat for the continued existence of humanity."[215] Geoffrey Hinton, known as one of the "godfathers of AI," left Google so that he could speak freely about his concern that AI poses an urgent risk to the survival of humanity.[216] Another AI pioneer, Yoshua Bengio, publicly claimed that "rogue AI may be dangerous for the whole of humanity."[217] In fairness, this is not a universal view. Yann LeCun, another pioneering figure, is famous for considering AI risk to be limited and to argue that the various risks will be worked out over time.[218]

Surveys among experts diverge considerably, although the average respondent sees a significant probability of a large-scale calamity. In one survey of AI and software engineers in Fortune 500 companies, the majority of respondents considered the possibility of (an undefined in time or scope) calamity from AI as higher than 25%.[219] Among the general public, a recent survey found that 9% of people believe that extinction risk is moderate or higher within the next ten years, and 22% see that level of risk over the next fifty years.[220] Another recent public survey found that 46% of respondents were "somewhat concerned" or more about the possibility of AI-caused extinction.[221] Among AI researchers, a 2022 survey found that the majority

---

215. Sam Altman, *Machine Intelligence, Part 1*, SAM ALTMAN BLOG (Feb. 25, 2015, 11:03 AM), https://blog.samaltman.com/machine-intelligence-part-1 [https://perma.cc/S7CL-YQBZ].

216. *See, e.g.*, Martin Coulter, *AI Pioneer Says Its Threat to World May Be 'More Urgent' Than Climate Change*, REUTERS (May 8, 2023, 11:19 PM), https://www.reuters.com/technology/ai-pioneer-says-its-threat-world-may-be-more-urgent-than-climate-change-2023-05-05/ [https://perma.cc/LJK4-UJ2P].

217. Bengio, *supra* note 214.

218. Yann LeCun (@ylecun), X (Apr. 2, 2023, 6:49 AM), https://x.com/ylecun/status/1642524629137760259 [https://perma.cc/7YGC-SH5Q].

219. *See* Barr Yaron, *State of AI Engineering 2023* (Oct. 9, 2023), https://elemental-croissant-32a.notion.site/State-of-AI-Engineering-2023-20c09dc1767f45988ee1f479b4a84135#694f89e86f9148cb855220ec05e9c631 [https://perma.cc/L8QT-7ZCM].

220. Jamie Elsey & David Moss, *US Public Opinion of AI Policy and Risk*, RETHINK PRIORITIES (May 12, 2023), https://rethinkpriorities.org/publications/us-public-opinion-of-ai-policy-and-risk [https://perma.cc/SJ9T-F8UM].

221. Taylor Orth & Carl Bialik, *AI Doomsday Worries Many Americans. So Does Apocalypse from Climate Change, Nukes, War, and More*, YOUGOV (Apr. 14, 2023, 2:16 PM), https://today.yougov.com/technology/articles/45565-ai-nuclear-weapons-world-war-humanity-poll [https://perma.cc/2DAK-CMVP].

of researchers believe that there is a 10% chance or more that AI will cause an existential catastrophe.[222] These surveys all ask different questions and follow different methodologies. Without putting too much stock in any single survey, the general picture is one where the possibility of large-scale harms from misaligned AI systems is receiving growing acceptance.[223] It is not universal, but it is no longer a fringe position.

In sum, we do not consider the likelihood of a large-scale AI calamity to be high, and an existential catastrophe is even less likely. But we do think there is enough theoretical and suggestive evidence that these risks must be taken seriously. We also note that, despite its importance, there has also been relatively little advancement in alignment theory and research.[224] Compared to the current explosion of investment in capabilities, the investment in safety and alignment is miniscule. We are hopeful that there is a solution, a set of solutions, or maybe just duct-taped kludges to the problem of alignment that are good enough. But as the technology currently stands, alignment is a major, unresolved concern.

III. THE CASE FOR SYSTEMIC REGULATION OF AI

The previous Part identified a variety of substantial, society-wide AI risks. Given the scope and magnitude of these risks, policymakers and other stakeholders should mitigate them, where feasible, either through regulation, informal guidance, or voluntary compliance. However, even accepting this basic premise, several questions remain. What form should AI risk mitigation take? Which risks should policymakers and others focus on? And, assuming regulation is appropriate, should lawmakers address these harms through targeted legislation, or should they regulate AI more systemically?

This Part addresses these questions. It contends that AI risk should be addressed largely through systemic regulation that governs AI as a technology, and that piecemeal laws will be insufficient to effectively regulate AI. It intervenes in ongoing debates about which potential AI harms deserve society's attention, arguing that viewing AI regulation as a zero-sum game is a mistake, and that recognition of both short- and long-term AI risk

---

222. Katja Grace et al., *2022 Expert Survey on Progress in AI*, AI IMPACTS (Aug. 3, 2022), https://aiimpacts.org/2022-expert-survey-on-progress-in-ai [https://perma.cc/UG4W-CYCN].

223. *See* sources cited *supra* note 27.

224. On the difficulties encountered by a well-funded organization, see Eliezer Yudkowsky, *MIRI Announces New "Death with Dignity" Strategy*, LESSWRONG (Apr. 1, 2022), https://www.lesswrong.com/posts/j9Q8bRmwCgXRYAgcJ/miri-announces-new-death-with-dignity-strategy [https://perma.cc/S6NP-W24M].

offers theoretical, practical, and political advantages. Finally, it addresses regulatory theory and the difficulties of cost-benefit analysis in the face of substantial uncertainty. It posits that, given the irreducible uncertainty of AI's future, a precautionary, maximin approach to regulation is justified.

A. Systemic AI Regulation

Addressing the AI risks discussed above will require government regulation. Private companies' voluntary compliance with industry guidelines may be sufficient in certain low-risk contexts[225] and could play a supportive role alongside legislative solutions. But, on its own, industry self-regulation would be woefully inadequate to address the society-wide risks of AI. These risks are largely inherent in the use of AI, and generally cannot be fixed through technical changes or the avoidance of obvious wrongdoing. Further, companies in a competitive market may have little incentive to use caution in AI development or deployment. Developing new AI capabilities and gaining a first-mover advantage over competing companies are such compelling economic goals for AI companies that compliance with voluntary industry guidelines is unlikely to be worthwhile.[226] Thus far, most AI companies have invested very little in AI safety research, instead devoting their resources to rapidly developing capabilities without regard to safety, transparency, or comprehension of how their systems operate.[227] Finally, past experience with industry self-regulation in various areas suggests that industry programs alone are unlikely to be effective, and are more likely to have a positive impact as complements to mandatory regulation.[228]

What form should AI regulation take? While issue-specific AI regulations will often be appropriate, more is needed to effectively address the society-wide risks of AI. Policymakers should regulate artificial intelligence systemically, as a technology, rather than solely on the basis of its applications. That is, as we describe below, meaningful AI regulation requires oversight of AI system development and deployment, rather than

---

225. *See infra* Section IV.C.

226. *See* Kolt, *supra* note 17.

227. Cristina Criddle & Madhumita Murgia, *Big Tech Companies Cut AI Ethics Staff, Raising Safety Concerns*, FIN. TIMES (Mar. 28, 2023), https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3.

228. *See, e.g.*, J. Alberto Aragón-Correa et al., *The Effects of Mandatory and Voluntary Regulatory Pressures on Firms' Environmental Strategies: A Review and Recommendations for Future Research*, 14 ACAD. MGMT. ANNALS 339, 339 (2020); Kendra Gray, *The Privacy Rule: Are We Being Deceived?*, 11 DEPAUL J. HEALTH CARE L. 89, 104–05 (2008).

particular AI applications alone.[229] It will require attention to system architecture, design, training, and testing, as well as use.[230]

Systemic regulation is necessary for several reasons. First, while some AI risks may be addressed by technical fixes or restrictions on obviously harmful or discriminatory uses, many AI risks are inherent in the technology itself.[231] Such intrinsic risks require a broader regulatory approach, because they exist wherever AI systems operate. Most of the potential harms detailed in Part II fit this description. As an example, using algorithms to sort people based on historical data inherently leads to discrimination. AIs that can infer the personal details of people's lives from their metadata threaten privacy by their very existence. Advanced AIs will pose threats to human employment by their very nature as systems capable of a wide variety of cognitive tasks. Highly capable and autonomous AIs would be dangerous because they are inherently unpredictable, difficult to understand, and extraordinarily powerful. These risks have to be mitigated at the development and design stages of the AI life cycle, as well as later stages.[232] In these contexts, regulators should determine whether and how AI systems can operate safely, not simply whether a system has caused some particular harm.

Second, the sheer number of risks posed by AI indicates that regulating AI as a technology will have substantial efficiency benefits over a piecemeal approach. Enacting separate laws to address each risk may be prohibitively difficult, costly, or time-consuming, or may leave obvious gaps. Systemic regulation requiring pre-approval of new AI systems can facilitate intervention at pre-deployment stages of AI development, addressing problematic or dangerous AI designs before they reach the public.[233] Moreover, systemic regulation can address both short and long-term risks in a comprehensive process. As explored further below, regulation targeting present AI harms can lay the groundwork for laws addressing novel or long-term risks, while addressing potential catastrophic harms can generate political and practical momentum for present-day legislation.[234]

Third, systemic regulation of AI systems is necessary because there is no guarantee that general purpose systems will only be used as intended by their developers. Containing AI systems once they are released can be difficult

---

229. *See infra* Section IV.A.

230. *See* Lehr & Ohm, *supra* note 19, at 655–57.

231. *See* Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1355–64 (2023) (discussing risks of AI, such as safety, employee recruitment, and public health); *supra* Section II.B; *infra* Section III.B.

232. Lehr & Ohm, *supra* note 19, at 655–57.

233. *See infra* Section IV.A.

234. *See infra* Sections III.B–C.

because they can be disseminated at low cost and their operation leaves little signature.[235] Already, after-market programmers have made their own connections between existing language models and various other software tools, creating, for example, a system meant to intentionally sow disinformation.[236] Because it will often be infeasible to regulate every downstream application of a system, it is critical to regulate the infrastructure itself. Relatedly, interventions at the research and development stage of machine learning models may be more effective and easier to design than those targeting deployed models.[237] Model design may also entail more human involvement and therefore greater transparency and more regulatory levers than post-development stages.[238]

Finally, new AI risks and harms may emerge over time, and they may be difficult to predict or prevent. Especially if AI capabilities continue to advance irregularly and at times sharply, regulators may struggle to keep up. Systemic approaches can help avert these novel harms without relying on policymakers to predict the future of AI. In this sense, systemically regulating AI systems can act as a catch-all for subtle or unrecognized AI harms. On their own, individualized approaches are brittle and porous, vulnerable to harms that are difficult to foresee.

Even establishing that AI will require systemic regulation leaves several foundational questions to be answered. There remains, for instance, the question of which AI harms policymakers should focus on when establishing systemic reviews of AI systems, and, indeed, which harms society should care about in conceptualizing AI risks.

B. Which Harms Deserve Our Attention?

From social media, to blogs, to op-ed pieces in major newspapers and academic journals, the debate over AI regulation has focused largely on a procedural question: should we focus our attention on the immediate harms of AI or the long-term risks that AI poses? Some writers focus on the possibility of AI superintelligence and threats of extinction, while ignoring

---

235. A popular language model, Bert, was downloaded 38 million times in February 2024 alone. *BERT Base Model (Uncased)*, HUGGING FACE, https://huggingface.co/bert-base-uncased [https://perma.cc/3N3W-NGHG]. While training large language models requires a large investment of compute resources, one can run a large language model on a consumer computer, leaving no signature.

236. *See* Pan et al., *supra* note 157.

237. *See* Lehr & Ohm, *supra* note 19, at 656–57 (explaining that the focus should be on the "playing with the data" stage because the "running-model stage" is too late).

238. *Id.* at 657.

harms caused by AI in the present day.[239] Sam Altman, the CEO of industry leader Open AI, takes this approach to its extreme, acknowledging the catastrophic risks of AI while lobbying against many forms of meaningful AI regulation in the short term.[240] Others take the opposite approach, arguing for an exclusive focus on immediate AI harms while dismissing concerns about long-term risks.[241] Some have even argued that experts' warnings about catastrophic AI risk will distract us from regulating AI in the present day.[242]

This debate, forged in the fires of Twitter feuds and online snark, has become counterproductive.[243] Working from mistaken premises about the zero-sum nature of AI concern, it presents a false choice. In reality, AI should be regulated because it causes immediate harms *and* threatens long-term catastrophe. Further, any political movement seeking meaningful AI regulation can only benefit from people recognizing both sets of potential AI harms. And many of the regulatory approaches that would effectively address short-term harms are appropriate first steps for regulating AI systems that threaten catastrophic harms.[244] Recognition of short-term and long-term AI risk is complementary, with each type of risk strengthening the case for meaningful regulation. We do not need to choose.

Regulating AI with a view towards immediate harms can lay the groundwork for future regulation of more dangerous AI. When initial AI regulations are in place, lawmakers can address new AI threats by amending existing laws rather than having to create new legislation from whole cloth. Litigation addressing immediate AI harms can bring malfunctioning systems to public attention before they cause widespread damage.[245] Laws may require government pre-screening for AI algorithms, giving regulators a

---

239. *See, e.g.*, Roman V. Yampolskiy, *Taxonomy of Pathways to Dangerous AI*, 2016 PROCS. 2D INT'L WORKSHOP ON AI, ETHICS & SOC'Y 143, https://arxiv.org/pdf/1511.03246.pdf [https://perma.cc/R2L4-YVBB] (discussing future risk of malevolent AI).

240. *See* Sam Altman et al., *Governance of Superintelligence*, OPENAI (May 22, 2023), https://openai.com/index/governance-of-superintelligence [https://perma.cc/G8TR-L2E9].

241. *See, e.g.*, Nir Eisikovits, *AI Is an Existential Threat—Just Not the Way You Think*, YAHOO! FINANCE (July 5, 2023), https://finance.yahoo.com/news/ai-existential-threat-just-not-122446498.html [https://perma.cc/CW9R-4T6C].

242. *Stop Talking About Tomorrow's AI Doomsday When AI Poses Risks Today*, 618 NATURE 885, 885 (2023).

243. Twitter is now "X," but the world still knows it as Twitter. Irina Ivanova, *Twitter Is Now X. Here's What That Means*, CBSNEWS (July 31, 2023, 5:18 PM), https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means/ [https://perma.cc/J953-U5UB].

244. *See infra* Sections IV.B–C.

245. *See infra* Section IV.B.

better chance to identify dangerous systems before they are deployed.[246] Other laws may deter development of open source or other hard-to-regulate forms of AI, reducing tortious practices and risky developmental approaches.[247]

On the other side, acknowledging the long-term catastrophic risks of AI can help justify systemic AI regulation in the present day. The costs and benefits of AI are uncertain, and so is AI's potential for catastrophic harm. But taking both short and long-term harm as real possibilities can help resolve any ambiguity regarding the appropriateness of regulation.[248] More practically, recognizing widespread concerns about catastrophic AI harms can bring attention, political momentum, and fundraising resources to the cause of AI regulation. It can motivate people and policymakers who may not normally be concerned about discrimination or privacy harms to support comprehensive AI regulation that can address those concerns. To build the largest and most effective coalition around AI regulation, it will be necessary to unify both sides of this argument in a single effort—one that recognizes all of the potential harms of AI, present and future.

We do not mean to argue that all AI regulation should be systemic, or that there are no worthwhile regulations that would only address immediate harms or long-term harms. Rather, we posit that (a) systemic regulation of AI is necessary and is an area of common ground between both camps in this debate, and (b) particularized AI regulations are also appropriate, but there is no reason to think that addressing one category of AI risk will impede addressing the other. Legislatures can pass laws specifically targeting AI discrimination or AI-based fraud, and also pass laws aimed at preventing self-improving AIs or the proliferation of autonomous weapons. A political culture that recognizes AI risk in one area is more likely to be open to recognizing it in another. By way of analogy, a polity that recognizes the long-term risks of climate change is also likely to recognize immediate climate change harms like extreme weather or environmental hazards—and vice-versa.[249] Identifying the issue and getting it on the policy agenda is the difficult step, and infighting among factions can only hinder that effort.

---

246. *See infra* notes 294–97 and accompanying text.

247. *See infra* notes 309–14 and accompanying text.

248. *See infra* Section III.C.

249. *See, e.g.*, Matthew T. Ballew et al., *Changing Minds About Global Warning: Vicarious Experience Predicts Self-Reported Opinion Change in the USA*, 173 CLIMACTIC CHANGE 1, 19 (2022) (reporting that experiencing or recognizing the impacts of climate change in the immediate term predicts changing one's opinion about climate change overall).

C. Costs, Benefits, and Catastrophic Harms

Artificial intelligence is a novel technology, already operating outside the realm of prior human experience. Its basic features distinguish it from prior technological breakthroughs.[250] Our previous technological advances—including technologies far more economically impactful than today's relatively limited AIs—could not write a sonnet, pass the Bar Exam, or draw a tree in a sunlit meadow. And AI's progress has been unpredictable and uneven, characterized by periods of minimal progress and sudden massive jumps in capabilities.[251] The future course of AI development is highly uncertain.

Under a standard cost-benefit approach to regulation, regulatory measures are justified when their benefits exceed their cost.[252] A starting point for assessing regulation of advanced technologies is the recognition that not every technological breakthrough results in a net positive outcome. For instance, germ-line gene editing, while promising, carries the potential to foster a form of genetic elitism and might inadvertently introduce unforeseen genetic disorders in subsequent generations.[253] Similarly, advancements in the synthesis of potent opioids—initially intended for pain relief—have fueled a public health crisis.[254]

It remains to be seen whether AI technology will be net positive or negative for society. We have detailed some of AI's potential risks above, but we also recognize the wide range of potential benefits. For example, some present and near-term benefits include improving agricultural yield;[255] enhancing environmental monitoring such as tracking deforestation and predicting natural disasters;[256] improving healthcare by offering personalized

---

250. *See supra* notes 23–26 and accompanying text.

251. *See supra* notes 171–77 and accompanying text; *supra* figs.1 & 3.

252. *See, e.g.*, David Parker & Colin Kirkpatrick, *Measuring Regulatory Performance*, ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT [OECD] 7 (2012), https://www.oecd.org/gov/regulatory-policy/3_Kirkpatrick%20Parker%20web.pdf [https://perma.cc/K5HE-MUNG] ("The critical public policy challenge is to ensure that the expected economic benefits from regulatory changes . . . outweigh any economic costs imposed.").

253. Eric Lander et al., *Adopt a Moratorium on Heritable Genome Editing*, 567 NATURE 165, 166–67 (2019).

254. *Addressing the Overdose Crisis*, U.S. DEP'T STATE, https://www.state.gov/addressing-the-overdose-crisis [https://perma.cc/648V-3HFP].

255. Qianyu Chen et al., *AI-Enhanced Soil Management and Smart Farming*, 38 SOIL USE & MGMT. 7, 8 (2022).

256. Jon Trask, *Harnessing the Power of AI and Blockchain to Combat Deforestation*, NASDAQ (June 23, 2023, 11:24 AM), https://www.nasdaq.com/articles/harnessing-the-power-

medicine;[257] early-diagnosis of disease, and cutting provision costs;[258] improving human access to information across language and cultural barriers;[259] optimizing education and training by creating personalized learning experiences;[260] improving energy efficiency by optimizing energy consumption;[261] offering more robust protection of human rights by improving monitoring of violations;[262] and improving disaster and disease response through improved prediction, logistics, and analysis.[263] Indeed, if we imagine highly capable AI systems, then this list is insufficiently ambitious. But even for moderately capable AI systems the benefits are likely to be broad and, in many cases, transformative.

Our aim is not to ban AI research and development. The focus should rather be on whether regulatory interventions are justified *on the margin*. And relative to the baseline of no meaningful regulation on AI systems (as opposed to specific application regulations),[264] there is a broad margin on which regulatory interventions are justified. As mentioned before, many of the potential upsides of AI necessarily entail large downsides. AI's potential of increasing of societal wealth would occur via massively displacing workers and dramatically increasing inequality.[265] AI's potential for efficient decision-making and prediction would also entail concretizing past

---

of-ai-and-blockchain-to-combat-deforestation [https://perma.cc/N4XT-KGBJ]; Monique M. Kuglitsch et al., *Facilitating Adoption of AI in Natural Disaster Management Through Collaboration*, 13 NATURE COMMC'NS 1, 1–2 (2022).

257. Agata Blasiak et al., *CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence*, 25 SLAS TECH. 95, 96 (2020).

258. Rebecca Fitzgerald et al., *The Future of Early Cancer Detection*, 28 NATURE MED. 666, 673 (2022).

259. Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. 83, 99–104 (2022).

260. Aditi Bhutoria, *Personalized Education and Artificial Intelligence in the United States, China, and India:* A Systematic Review Using a Human-in-the-Loop Model, 3 COMPUTS. & EDUC.: A.I. 1, 2 (2022).

261. Yassine Himeur et al., *Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives*, 287 APPLIED ENERGY 1, 2 (2021).

262. Nenad Tomašev et al., *AI for Social Good: Unlocking the Opportunity for Positive Impact*, 11 NATURE COMMC'NS 1, 3–4 (2020).

263. Wenjuan Sun et al., *Applications of Artificial Intelligence for Disaster Management*, 103 NAT. HAZARDS 2631, 2632 (2020).

264. The FTC has issued relevant guidance in the context of credit decisions. *See* Andrew Smith, *Using Artificial Intelligence and Algorithms*, FED. TRADE COMM'N: BUS. BLOG (Apr. 8, 2020), https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-and-algorithms [https://perma.cc/8UJ7-RGXF].

265. *See supra* Section I.B.1.

discrimination and violating consumer privacy in unprecedented ways.[266] Improvements in facial recognition and other AI surveillance technologies can increase security and law enforcement productivity, but would decrease citizen autonomy and liberty.[267] Automated AI weapons reduce troop casualties and create more effective weapons of war, but also lower the cost of starting conflicts, create serious risks of misalignment, and increase the likelihood of imperialism and totalitarianism.[268] There are also downsides with no corresponding upside, including enhanced fraud and scams, more effective terrorism, and greater quantities of misinformation.[269]

In this sense, AI systems belong to a large family of technologies that, while beneficial, pose substantial risks of harm and require regulation. Burning coal for power has been extremely beneficial historically, especially for developing nations.[270] Nuclear power can efficiently provide energy, free of carbon emissions.[271] Research on deadly viruses can lead to new vaccines and treatments.[272] But each of these beneficial technologies is also extremely dangerous if left unregulated. We do not allow just anyone to operate a nuclear reactor or use deadly viruses for research, and we increasingly regulate the burning of fossil fuels, because of these dangers.[273] Even with a very optimistic view of AI's harms and benefits, there is ample reason to support regulation.

In assessing potential AI regulation, we need to be aware of both the individual and the societal risks that AI entails. We cannot tell now what the net effect will be, but the balance will surely be higher if the negative outcomes can be avoided. Moreover, the non-trivial risk of mass calamities

---

266. *See supra* Sections I.A.1, I.A.3.

267. *See* Selinger & Hartzog, *supra* note 89, at 111.

268. *See supra* Sections I.B.2–3.

269. *See supra* Sections I.A.2, I.B.3–4, II.A–B.

270. *See, e.g.*, Samantha Gross, *Why Are Fossil Fuels So Hard to Quit?*, BROOKINGS INST. (June 2020), https://www.brookings.edu/articles/why-are-fossil-fuels-so-hard-to-quit [https://perma.cc/5JJY-U8LD].

271. *See, e.g.*, Thomas E. Rehm, *Advanced Nuclear Energy: The Safest and Most Renewable Clean Energy*, 39 CURRENT OP. CHEM. ENG'G 1, 1 (2023).

272. Andy Kilianski et al., *Gain-of-Function Research and the Relevance to Clinical Practice*, 213 J. INFECTIOUS DISEASES 1364, 1367 (2016).

273. *See, e.g.*, *Nuclear Power Plant Licensing Process*, U.S. NUCLEAR REGUL. COMM'N (July 2009), https://www.nrc.gov/reading-rm/doc-collections/nuregs/brochures/br0298/index.html [https://perma.cc/FYL9-PP88]; *Gain of Function Research*, NAT'L INSTS. OF HEALTH, https://osp.od.nih.gov/policies/national-science-advisory-board-for-biosecurity-nsabb/gain-of-function-research [https://perma.cc/BP42-R9R4] (last updated Apr. 2023); Camila Domonoske, *The Big Reason Why the U.S. Is Seeking the Toughest-Ever Rules for Vehicle Emissions*, NPR (Apr. 12, 2023, 5:01 AM), https://www.npr.org/2023/04/12/1169269936/electric-vehicles-emission-standards-tailpipes-fuel-economy [https://perma.cc/6VF9-J9YS].

that AI poses, identified by countless experts,[274] must be included in an accurate cost-benefit analysis of AI development.

There is an additional argument for AI regulation that rests on the deep uncertainty surrounding its future development. Regulation skeptics may argue that because we cannot predict AI's risks with certainty, we should be skeptical that they will ever arise. Yet AI's future benefits are equally uncertain and probabilistic. There is, at heart, an irreducible degree of uncertainty on both sides of the ledger.

In situations of probabilistic uncertainty, precautionary regulatory approaches may be justified.[275] This is especially the case when the thing to be regulated creates a non-trivial risk of catastrophic harm.[276] As Sunstein notes, the very idea of the "Precautionary Principle might well be reformulated as an Anti-Catastrophe Principle, designed for special circumstances in which it is not possible to assign probabilities to potentially catastrophic risks."[277] For example, governments may be justified in precautionary regulation of pollutants that cause climate change, because the effects of climate change are uncertain and its downside risks are potentially catastrophic.[278] Even Richard Posner concludes that for uncertain large scale catastrophes, "it behooves us to give serious consideration to increasing our efforts at prevention."[279]

A notable precautionary approach involves the pursuit of a *maximin* strategy. Under this strategy, the way to deal with uncertain futures is by choosing the policy approach with the best worst-case outcome.[280] Regulators should attempt to prevent plausible worst-case scenarios rather than waiting years or decades for probabilistic uncertainty to resolve.[281] Such a strategy may maximize welfare in situations of uncertainty and substantial potential harms.[282]

---

274. *See* sources cited *supra* note 27.

275. *See, e.g.*, Cass R. Sunstein, *Maximin*, 37 YALE J. ON REGUL. 940, 967 (2020); JOHN RAWLS, A THEORY OF JUSTICE 132–39 (1999); JON ELSTER, EXPLAINING TECHNICAL CHANGE: A CASE STUDY IN THE PHILOSOPHY OF SCIENCE 186–207 (1983).

276. Sunstein, *supra* note 275, at 966.

277. *See* CASS R. SUNSTEIN, LAWS OF FEAR: BEYOND THE PRECAUTIONARY PRINCIPLE 5 (2005).

278. *See* STEPHEN M. GARDINER, A PERFECT MORAL STORM: THE ETHICAL TRAGEDY OF CLIMATE CHANGE 411–14 (2011).

279. RICHARD POSNER, CATASTROPHE: RISK AND RESPONSE 198 (2004). Posner contemplates bioterrorist attacks, but his argument is not specific to this type of risk. *Id.*

280. *See* Sunstein, *supra* note 275, at 943, 965–66.

281. *See id.*

282. *See id.* at 976.

Artificial intelligence is precisely the type of technology for which a maximin, precautionary regulatory strategy is appropriate. The path of its future development is uncertain, and, according to hundreds of experts in the field of AI development, it poses a substantial risk of catastrophic harm.[283] To be sure, some would argue that we should charge ahead because AI's benefits will eclipse its risks and a maximin strategy would needlessly prevent us from realizing those large benefits.[284]

Yet these arguments are flawed, for at least four reasons. First, as noted above, many of the more plausible benefits of AI (economic growth, efficient algorithmic prediction) inherently carry with them substantial harms (inequality and joblessness, discrimination, and privacy invasions).[285] Moreover, regulation does not have to prevent any and all AI deployment. A regulatory regime does not mean a complete ban.

Second, even if AIs are far more likely to bestow miraculous benefits on humanity than it currently appears, maximin strategies are often appropriate to prevent large catastrophes even at the expense of preventing massive gains.[286] For example, precautionarily avoiding extinction may be justified even if the foregone upsides are enormous, in part because human existence is already extremely valuable and because humans are likely to continue to innovate even without the assistance of super-capable AIs.

Third, AI regulation can be flexible in response to extraordinary circumstances. It is possible that strong AI systems may someday help address threats of extinction, like a hurtling asteroid or an exceptionally lethal pandemic.[287] Yet this distant possibility need not undermine the case for AI regulation. If such risks ever become real, the regulatory apparatus could be relaxed and scaled down as an emergency measure, until the threat is

---

283. *See* sources cited *supra* note 27.

284. *See, e.g.*, David Streitfeld, *Silicon Valley Confronts the Idea That the 'Singularity' Is Here*, N.Y. TIMES (June 11, 2023), https://www.nytimes.com/2023/06/11/technology/silicon-valley-confronts-the-idea-that-the-singularity-is-here.html; Hasan Chowdhury, *Get the Lowdown on 'e/acc'—Silicon Valley's Favorite Obscure Theory About Progress at All Costs, Which Has Been Embraced by Marc Andreessen*, BUS. INSIDER (July 28, 2023, 6:44 AM), https://www.businessinsider.com/silicon-valley-tech-leaders-accelerationism-eacc-twitter-profiles-2023-7 [https://perma.cc/27PJ-PAR7].

285. *See supra* Part II.

286. Sunstein, *supra* note 275, at 964–65.

287. *See, e.g.*, Robert Lea, *AI Algorithm Discovers 'Potentially Hazardous' Asteroid 600 Feet Wide in a 1st for Astronomy*, SPACE.COM (Aug. 8, 2023), https://www.space.com/ai-finds-first-potentially-dangerous-asteroid [https://perma.cc/63D7-9ZL5].

resolved. With such an approach, the prevention of AI mass risk could co-exist to some degree with AI protection from mass risks.[288]

Finally, we think there is a prima facie ethical duty to err on the side of caution. Even if the chances of a miraculous future are higher than the chances of extinction, morality and pragmatism may dictate that we take the safer route. That is, as discussed further below, we may have a moral duty to avoid significant extinction risks and preserve humanity, even if doing so requires foregoing considerable benefits.[289] This is especially true since speeding up will remain an option for future generations, if they deem the calculus to have sufficiently changed. But given current epistemic uncertainties, we think there is a moral command to treat humanity with the dignity it deserves.

Human extinction, were it to occur in the next century, would result in the deaths of every person then living—billions or tens of billions of deaths. This would be a horror on a scale beyond our comprehension, the equivalent of every death experienced in the worldwide COVID-19 pandemic occurring in a single hour, and then a second pandemic occurring again the next hour, and then a third occurring the next hour, and a fourth, and a fifth, every hour, for months, until everyone was gone.[290] Yet total extinction would be a harm far greater than the immense sum of this loss. It would be the end of humanity, and all that humanity means.

Much of the lasting significance of our lives resides in our contributions, however small, to the broader narrative of human existence. Our actions have some meaning and impact even after our deaths because they help shape the future of humanity in its ongoing struggle to survive and flourish in a vast, indifferent universe.[291] Extinction ends that struggle and erases that meaning. More broadly, extinction ends the human narrative before it fully develops, confining humanity's existence to a far narrower block of time than most species experience and curtailing all the good that humanity might someday

---

288. The critic may then retreat to the position that regulation would stall innovation such that when imminent threats are discovered, scaling down regulation would not allow enough time for development of effective solutions. But this argument cannot justify, in our view, avoiding all regulation against known and unknown risks simply to gain marginal increase in preparedness against uncertain risks.

289. *See, e.g.*, BRIAN GREENE, UNTIL THE END OF TIME: MIND, MATTER, AND OUR SEARCH FOR MEANING IN AN EVOLVING UNIVERSE 319 (2020); SAMUEL SCHEFFLER, DEATH AND THE AFTERLIFE 59–60 (Niko Kolodny ed., 2013).

290. *See WHO COVID-19 Dashboard*, WORLD HEALTH ORG., https://covid19.who.int [https://perma.cc/7XYU-SL8J].

291. *See, e.g.*, Ward & King, *supra* note 120, at 61; Costin & Vignoles, *supra* note 121, at 865.

do. A significant part of all the sacrifices made and work done for the betterment of humanity—the noblest instances of human achievement and charity—will have been in vain.[292] Regulating new technologies to address non-trivial threats of extinction is, in short, amply justified.

IV. TOWARDS SYSTEMIC AI REGULATION

How should we approach the risks and challenges discussed above? This Part addresses that question. The possibilities for AI regulation in the United States are broad and varied. But while U.S. policymakers have begun the process of gathering information about the topic, much of the conceptual work necessary for substantive AI regulation against broad societal risks remains to be done.[293] In this Part, we begin that work.

A. Domestic Regulation

This Section's focus is on general principles of AI regulation, rather than particular proposals or draft legislation. Nonetheless, our proposed principles are more concrete and pragmatic than prior efforts in the early theoretical literature on comprehensive AI regulation.[294] The principles are intended to move society closer to meaningful AI governance by providing both clear guidance and a variety of options to policymakers. We set them out below.

First, AI regulation should be *systemic*, regulating artificial intelligence as a technology rather than solely on the basis of its applications. In a recent congressional hearing, an IBM representative insisted that Congress should only regulate AI applications, such as when an AI system is involved in making credit decisions or screening job applicants.[295] This is a myopic approach. For all of the reasons discussed above, the society-wide risks of AI will require systemic regulation to effectively address.

Second, and relatedly, effective AI regulation will require *ex ante oversight and approval of AI system development and deployment*. Ex post regulation via government or private enforcement, while a potentially valuable part of a regulatory regime, is insufficient on its own to successfully regulate AI. Courts are likely to be overworked and underresourced; AI harms will often be difficult to identify or trace to a specific wrongdoer;

---

292. *See* sources cited *supra* note 289.

293. *See* sources cited *supra* note 21.

294. *See* Kolt, *supra* note 17; Chesterman, *supra* note 17.

295. *See AI Hearing*, *supra* note 40, at 3–5 (statement of Christina Montgomery, Chief Privacy and Trust Officer, IBM).

enforcement may be slow even once the responsible party is identified; and penalties may be insufficient to deter wrongdoing.[296] Instead, ex ante review of AI systems and applications is likely necessary to prevent serious harms. Many harms could be mitigated through regulatory interventions at the design and development stages, requiring, for example, the inclusion of best alignment practices in the training of the system, or the exclusion of elements that could give the system control of the reporting of its training progress.[297]

Here too, ex ante oversight should be systemic. Regulation should cover system architecture, design of system objectives, training runs, testing, and finally, deployment. At any one of these stages, critical errors may emerge that might be unfixable in hindsight. The experience of OpenAI, in which a training run was accidentally set to maximize human disapproval (because they multiplied the objective by -1), should be treated as a major accident.[298] Preventing the creation or deployment of dangerous AI systems is far more effective, and likely far more efficient, than attempting to address them once they are in use.

More broadly, a licensing regime for AI could require firms to secure regulatory pre-approval before developing a new AI system or applying an AI in a new context. This may require providing sufficient justifications along several dimensions including safety, nondiscrimination, accuracy, transparency, accountability, scenario planning, and/or resilience in the event of disaster, depending on the system at issue.[299] Licensing can also ensure that firms maintain and update AIs that play critical roles in decision-making, transportation, or other important contexts.[300] Finally, licensure can allow

---

296. Gianclaudio Malgieri & Frank Pasquale, *Licensing High-Risk Artificial Intelligence: Toward Ex Ante Justification for a Disruptive Technology*, 52 COMPUT. L. & SEC. REV. 1, 1–2 (2024), https://www.sciencedirect.com/science/article/pii/S0267364923001097?ref=pdf_download&fr=RR-2&rr=8636709ffc55a6ee [https://perma.cc/T93E-8YC2].

297. *See* Tutt, *supra* note 41, at 117.

298. Daniel M. Ziegler et al., Fine-Tuning Language Models from Human Preferences (Jan. 8, 2020) (unpublished manuscript), https://arxiv.org/pdf/1909.08593 [https://perma.cc/2U2C-DZZA] ("One of our code refactors introduced a bug which flipped the sign of the reward. . . . The result was a model which optimized for negative sentiment while still regularizing towards natural language. Since our instructions told humans to give very low ratings to continuations with sexually explicit text, the model quickly learned to output only content of this form. This bug was remarkable since the result was not gibberish but maximally bad output. The authors were asleep during the training process, so the problem was noticed only once training had finished.").

299. *E.g.*, Tutt, *supra* note 41, at 116–17; Malgieri & Pasquale, *supra* note 296, at 1–2, 9–11.

300. *See* Malgieri & Pasquale, *supra* note 296, at 3.

policymakers to permit high-value, low-risk uses of AI while prohibiting more dangerous or less beneficial applications.[301]

Third, domestic AI regulation should be *strategically compatible with, but independent of, international regulation*. Domestic policymakers may be reluctant to restrain the local AI industry to a vastly greater extent than other countries. They might fear that such regulation will place the United States at an economic or military disadvantage.[302] We agree that effective regulation will require international cooperation, and we return to this point below. But we also think it would be unwise for the United States, which is a leader in the field, to drag its feet in face of substantial AI risks.

There is room for significant domestic AI regulation even in the absence of international action. Currently, cutting-edge AI research is largely concentrated in the United States and China, and to a lesser extent Europe.[303] Thus far, China and the European Union have been substantially more active in regulating AI development than the United States.[304] These countries' laws are discussed further in Section IV.C. Their approaches might provide a partial template for early-stage AI regulation in the United States, although the U.S. should aspire to recognize broader categories of risk.[305] Domestic legislation can additionally facilitate international cooperation by signaling a genuine commitment to regulating AI.

In the short term, the United States might also pass laws restricting investments in foreign AI companies, or perhaps impose curbs on international sales of the U.S.-produced microchips used in cutting-edge AI data centers in addition to those the Biden administration enacted in October 2022.[306] Alternatively, it might adopt a more cooperative policy and fewer hardware restrictions. Whatever the approach, domestic legislation should harmonize with the United States' international AI strategy.

---

301. *See id.* at 15.

302. *See, e.g.*, sources cited *infra* note 355; Amanda Askell et al., The Role of Cooperation in Responsible AI Development (July 10, 2019) (unpublished manuscript), https://arxiv.org/pdf/1907.04534.pdf [https://perma.cc/E9KA-7VTD].

303. *See, e.g.*, Neil Savage, *Learning the Algorithms of Power*, 588 NATURE S102, S102–03 (2020).

304. *See infra* Section IV.C.5.

305. *See infra* notes 409–30 and accompanying text.

306. Ana Swanson et al., *Biden Administration Weighs Further Curbs on Sales of A.I. Chips to China*, N.Y. TIMES (June 28, 2023), https://www.nytimes.com/2023/06/28/business/economy/biden-administration-ai-chips-china.html; *see also* Ben Wodecki, *Biden Targets Chinese AI Development with Potential Cloud Service Ban*, AI BUS. (July 5, 2023), https://aibusiness.com/verticals/biden-targets-chinese-ai-development-with-potential-cloud-service-ban [https://perma.cc/VB7N-PPUF].

Fourth, regulatory efforts should *promote and incentivize alignment research*. While market participants have a natural incentive to invest in capabilities development, they have considerably less incentive to invest in making sure their products are safe and aligned.[307] Currently, research on alignment is poorly organized. For example, there are many professors studying AI, but few that specialize in alignment per se. Governments should invest in foundational alignment research, for instance via generous research grants and subsidies. But AI companies, where knowledge of development and safety issues is concentrated, should also play an active role in such research. To prevent companies from neglecting AI safety in their race for market share, legislation could require that companies developing AI capabilities also invest significant resources in alignment research.[308]

Fifth, AI regulation should *employ a diverse set of regulatory approaches*. AI presents a wide array of potential harms, some of which are extraordinarily dangerous. Employing a variety of procedures for AI regulation can help address this broad range of harms and ensure that the failure of one set of measures does not lead to catastrophe.[309] The causes of AI harm are also complex and can arise at different stages of AI development.[310] In the face of deep uncertainty, policymakers should use a variety of regulatory tools that target the many stages of the AI process.[311]

Sixth, AI regulation should *address, at the very least, the most obvious pathways to harm or catastrophe.* Some AI applications are primarily useful for facilitating fraud or tortious activity. For instance, voice cloning services are now widely available, and customers can clone the voices of others as well as their own.[312] Deepfake generators can help users create realistic fake videos based on existing videos of virtually anyone they choose.[313] While technologies like this do have some non-harmful uses—perhaps gaming and movie production—they are easily deployed as scalable tools for engaging in

---

307. *See* Kolt, *supra* note 17.

308. Part of the alignment effort should also be directed toward public dissemination of information on the successes and failures of AI safety. We also see a role for government-organized research, such as that conducted by the RAND Corporation, which would focus on broad, foundational work.

309. Kolt, *supra* note 17.

310. *Id.* at 47.

311. *Id.*

312. *See* sources cited *supra* note 71.

313. *See* Ceclia Hwung, *How to Make a DeepFake Video*, DIGIARTY, https://www.videoproc.com/video-editor/how-to-make-a-deepfake-video.htm [https://perma.cc/C5HJ-B9LH] (Apr. 29, 2024).

fraudulent, tortious, harassing, or discriminatory behavior.[314] Technologies like this are ripe targets for regulation or prohibition.

Similarly, some AI development practices may be especially reckless or closely associated with potential downside risks. Recursively self-improving AIs, AIs that modify their own source code, highly autonomous AIs, and AI systems that are connected to a broad array of physical tools are especially likely to develop alignment problems or dangerous capabilities of the type that raise concerns about catastrophic risks.[315] Attempts to develop such AIs are particularly well-suited to precautionary regulation or prohibition. And while none of these AIs has yet been deployed in its full form, developers have created preliminary versions, with AIs that create detailed code, AIs that recursively generate questions to ask themselves in order to efficiently complete a task, and AIs that conduct internet research and use what they learn to complete tasks.[316]

Regulators should also develop a cautious approach to open sourcing of AI models. Smaller, vetted systems may well contribute to experimentation and alignment efforts by individuals or small groups. But the broad sharing of models has already proven itself problematic, with users fine-tuning large models on the toxic and racist content of 4Chan, models trained to create malware, and models that specialize in spam and disinformation generation.[317] Private individuals have connected AIs to a variety of tools,

---

314. *See* Carter Evans & Analisa Novak, *Scammers Use AI to Mimic Voices of Loved Ones in Distress*, CBS NEWS (July 19, 2023, 9:48 AM), https://www.cbsnews.com/news/scammers-ai-mimic-voices-loved-ones-in-distress [https://perma.cc/5U43-VA7G].

315. *See* Kolt, *supra* note 17, at 1192–93.

316. *See, e.g.*, Mark Sullivan, *Auto-GPT and BabyAGI: How 'Autonomous Agents' Are Bringing Generative AI to the Masses*, FAST CO. (Apr. 13, 2023), https://www.fastcompany.com/90880294/auto-gpt-and-babyagi-how-autonomous-agents-are-bringing-generative-ai-to-the-masses [https://perma.cc/SV6J-TWVQ]; Tanya Malhotra, *Breaking Down AutoGPT: What It Is, Its Features, Limitations, Artificial General Intelligence (AGI) and Impact of Autonomous Agents on Generative AI*, MARKTECHPOST (July 11, 2023), https://www.marktechpost.com/2023/07/11/breaking-down-autogpt-what-it-is-its-features-limitations-artificial-general-intelligence-agi-and-impact-of-autonomous-agents-on-generative-ai/ [https://perma.cc/Z43L-ZXPX].

317. *See, e.g.*, Tianle Cai et al., Large Language Models as Tool Makers (May 26, 2023) (unpublished manuscript), https://arxiv.org/pdf/2305.17126.pdf [https://perma.cc/X3VX-9EWH]; Pan et al., *supra* note 157; Xiangyu Qi et al., Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! (Oct. 5, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.03693.pdf [https://perma.cc/7J6Q-RNCA]; Stuart A. Thompson, *Dark Corners of the Web Offer a Glimpse at A.I.'s Nefarious Future*, N.Y. TIMES (Jan. 8, 2024), https://www.nytimes.com/2024/01/08/technology/ai-4chan-online-harassment.html.

and the process is largely irreversible.[318] Restrictions on public dissemination of AI architecture, weights, biases, and even some forms of output may help prevent serious harms.

Finally, AI regulation *can benefit from state as well as federal involvement*. States can adopt a variety of legislative approaches, and other states, the federal government, and foreign governments can learn from their successes and failures. AI regulation may be especially likely to benefit from states' experimenting with a wide range of new approaches.[319] In recent years, state legislatures have usefully regulated harmful AI applications in the absence of federal legislation.[320] For example, several states and cities have recently banned forms of AI-driven surveillance, offering their citizens substantial protections.[321] Even after the federal government has regulated a new technology, states may be able to enact additional restrictions on it without being preempted, depending on the character of the state restriction and the specifics of the federal law.[322] State policymakers should inform themselves about AI risks and benefits and move forward with AI regulation, consistent with the principles discussed here.

B. Litigation

Courts and litigants have an important role to play in regulating artificial intelligence. AIs, and entities using AI, will inevitably commit various torts and other civil violations—indeed they have already done so.[323] Civil

---

318. JAMES BRIGGS & FRANCISCO INGHAM, LANGCHAIN AI HANDBOOK chs. 5–6 (n.d.), https://www.pinecone.io/learn/series/langchain/.

319. *See supra* text accompanying notes 309–11.

320. *See, e.g.*, Brenna Goth, *Illinois 'Deepfake' Law Penalizes Sharing Altered Sexual Images*, BLOOMBERG L. (July 28, 2023, 2:31 PM), https://news.bloomberglaw.com/ip-law/illinois-deepfake-law-penalizes-sharing-altered-sexual-images; Geoff Mulvihill, *What to Know About How Lawmakers Are Addressing Deepfakes like the Ones that Victimized Taylor Swift*, ASSOCIATED PRESS (Jan. 31, 2024), https://apnews.com/article/deepfake-images-taylor-swift-state-legislation-bffbc274dd178ab054426ee7d691df7e [https://perma.cc/T5YW-2A47].

321. *See, e.g.*, Grace Woodruff, *Maine Now Has the Toughest Facial Recognition Restrictions in the U.S.*, SLATE (July 2, 2021, 5:50 AM), https://slate.com/technology/2021/07/maine-facial-recognition-government-use-law.html [https://perma.cc/M6TE-YKYC]; *Vermont Lawmakers Approve Ban on Facial Recognition Technology*, WCAX (Oct. 13, 2020, 3:51 PM), https://www.wcax.com/2020/10/13/vermont-lawmakers-approve-ban-on-facial-recognition-technology [https://perma.cc/68US-DZ9T].

322. *See* Doug Farquhar & Liz Meyer, *State Authority to Regulate Biotechnology Under the Federal Coordinated Framework*, 12 DRAKE J. AGRIC. L. 439, 461–72 (2007).

323. *See* Bryan Pietsch, *2 Killed in Driverless Tesla Car Crash, Officials Say*, N.Y. TIMES (Nov. 10, 2021), https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html;

litigation can compensate plaintiffs for AI harms from physical injuries to privacy invasions, medical errors, civil rights violations, fraud, manipulation, and more.[324] Constitutional litigation involving unlawful discrimination claims may provide important deterrence against bias in algorithmic decision-making.[325] Finally, intellectual property infringement claims could bring useful judicial scrutiny to the training practices of AI developers, which often involve the processing of copyrighted or otherwise protected works.[326]

Establishing a clear doctrinal path for persons harmed by AIs to bring civil claims can also contribute toward effective systemic regulation of AI. Lawsuits can act as an early warning system for dangerous or poorly designed AIs. When an AI system causes harm, an injured person should not be limited to petitioning the government and hoping it eventually addresses the issue. Filing a lawsuit brings the problem to public notice more quickly than lobbying for government action typically would, and courts can generally respond to harms long before legislatures do.[327]

Further, litigation can act as a regulatory tool in its own right, providing incentives to developers to carefully assess the risks and benefits of their AIs rather than hastily deploying potentially dangerous systems.[328] Liability can motivate developers to pre-test AI performance, bolster data security, gather information about how their AIs operate, and take other safety-improving steps that they might otherwise skip in order to hasten their products to market.[329]

Attorneys and judges can draw on a rich existing literature of helpful proposals for applying traditional forms of liability to the novel context of AI actors. To illustrate, in torts, many scholars have argued in favor of a strict

---

Neal E. Boudette, *Tesla's Autopilot Technology Faces Fresh Scrutiny*, N.Y. TIMES (Mar. 23, 2021), https://www.nytimes.com/2021/03/23/business/teslas-autopilot-safety-investigations.html.

324. *See, e.g.*, Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1319–20 (2020); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 902 (2017).

325. *See, e.g.*, Emily Black et al., *Less Discriminatory Algorithms*, 113 GEO. L.J. (forthcoming 2024); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 291 (2020).

326. *See, e.g.*, Lemley & Casey, *supra* note 18, at 746–48.

327. *See, e.g.*, Matthew Tokson, *Knowledge and Fourth Amendment Privacy*, 111 NW. U. L. REV. 139, 193 (2016).

328. *See* Omri Rachum-Twaig, *Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots*, 2020 U. ILL. L. REV. 1141, 1163–64 (2019).

329. *See id.*

liability approach for harms caused by AI systems.[330] They contend that AI developers are in a better position to anticipate and prevent risk and that proof is likely especially challenging in these scenarios.[331] Others have suggested applying this framework to securities violations by trading algorithms and antitrust violations when algorithms unlawfully collude.[332]

We close with one cautionary note. Litigation can reveal *too much* information. We consider information about specific model architecture, training techniques, certain benchmark results, and even some model outputs as sensitive information. Courts should be extremely cautious about inclusion of this information in public filings.[333] In certain cases, in camera review will be appropriate.

C. International Governance

Effective governance of AI will require an international component. Large AI systems reside in computing centers that often cross political boundaries.[334] In a globalized world, the harms from AI systems will not be contained to a single country, and several more extreme forms of harm may well endanger global order or human existence altogether. An international response is necessary.

But is it possible? If AI promises power, nation-states may rush to develop it for themselves, because even if they themselves understand the danger, their rivals might be less careful. This could jumpstart a race to the bottom,

---

330. *See, e.g.*, Abraham & Rabin, *supra* note 18, at 153–54; David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 146–47 (2014).

331. *See* Rachum-Twaig, *supra* note 328, at 1162–64.

332. Diamantis, *supra* note 18, at 801–05; Greg Rosalsky, *When Computers Collude*, NPR: PLANET MONEY (Apr. 2, 2019), https://www.npr.org/sections/money/2019/04/02/708876202/when-computers-collude [https://perma.cc/V8BY-EKWC].

333. *See* Gregory Gerard Greer, *Artificial Intelligence and Trade Secret Law*, 21 U. ILL. CHI. REV. INTELL. PROP. L. 252, 264–65 (2022); Sumeet Wadhwani, *Open Source vs. Proprietary AI: A Tussle for the Future of Artificial Intelligence*, SPICEWORKS (Dec. 12, 2023), https://www.spiceworks.com/tech/artificial-intelligence/articles/open-source-vs-proprietary-ai-development/ [https://perma.cc/Q5R9-C7E8]; *cf.* Omri Ben-Shahar & Lisa Bernstein, *The Secrecy Interest in Contract Law*, 109 YALE L.J. 1885 (2000).

334. Michael Veale et al., *AI and Global Governance: Modalities, Rationales, Tensions*, 19 ANN. REV. L. & SOC. SCI. 255, 265 (2023); Effy Vayena & Andrew Morris, *A Bioethicist and a Professor of Medicine on Regulating AI in Health Care*, ECONOMIST (Feb. 28, 2023), https://www.economist.com/by-invitation/2023/02/28/a-bioethicist-and-a-professor-of-medicine-on-regulating-ai-in-health-care.

where even responsible nations will feel pressure to charge ahead without sufficient safeguards.

Fortunately, history provides some positive guidance. AI is not the first technology to provide military and economic advantages while imposing serious risks.[335] Yet there are several precedents of nations avoiding vicious dynamics through governance and collaboration.[336] From the laws of just war to limits on pollution, and from physics research to investment in international measures against pandemics, nation-states are capable of avoiding races to the bottom and enabling effective joint action.

There is also an interesting dynamic between our discussion in the prior sections and the current one. Many successful international measures emerge from effective domestic regulation, and then inspire further domestic regulation.[337] Our goal here is to explore the various lessons from international law for the problem of regulating AI.

The following discussion considers several possible modes of international governance for AI: transparency & opacity mechanisms, harmonization measures, technology assessment, soft law, and hard law. These modes represent a range of AI oversight options that are neither mutually exclusive nor exhaustive.

1. Transparency & Opacity

Effective regulation of AI technology involves a smart mix of transparency and opacity measures. Transparency is positive when it promotes alignment research, enables effective monitoring of investments in potentially dangerous capabilities, and facilitates accountability among decisionmakers if they are too lax with regulated firms. Transparency is risky when it discloses machine learning techniques and architectures; when it

---

335. KELLEY SAYLER, CONG. RSCH. SERV., R46458, EMERGING MILITARY TECHNOLOGIES: BACKGROUND AND ISSUES FOR CONGRESS 1 (2024), https://sgp.fas.org/crs/natsec/R46458.pdf [https://perma.cc/UPA5-QYCA].

336. *See, e.g.*, Martyn P. Chipperfield et al., *Quantifying the Ozone and Ultraviolet Benefits Already Achieved by the Montreal Protocol*, NATURE COMMC'NS (May 26, 2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4455099 [https://perma.cc/R7SJ-8MEK] (discussing progress in restoring the ozone layer through the Montreal Protocol and subsequent amendments and adjustments); Glenn Cross & Lynn Klotz, *Twenty-First Century Perspectives on the Biological Weapon Convention: Continued Relevance or Toothless Paper Tiger*, 76 BULL. ATOMIC SCIENTISTS 185, 185 (2020) (recounting how the Biological Weapons Convention "has successfully bolstered the near universal norms against the use of biological weapons").

337. *Transparency and Explainability (Principle 1.3)*, ORG. FOR ECON. COOP. & DEV., https://oecd.ai/en/dashboards/ai-principles/P7 [https://perma.cc/W2HH-EJSE].

reveals information that might jumpstart new lines of capability research; and even when it leaks model outputs that can later be reverse-engineered. The problem is complex, and a pluralistic regime is appropriate.

The goal of transparency incorporates a number of values. One set of issues, recognized by the OECD AI group, relates to explainability.[338] Here, transparency can play a role in mitigating bias and increasing comprehension of AI operations.[339] Transparency can also be used to track significant developers, infrastructure providers, and related players—so that if concerns emerge, these actors will be easier to hold to account. Another goal of transparency consists of sharing ideas and strategies on alignment and safety with the larger research community.[340] Governments should be made aware if models, anywhere in the world, engage in unwanted behavior, including lab accidents, attempts to copy themselves, or instances of deceit.

One promising method of tracking development is public registries. Public registries are an important transparency mechanism for the governance of emerging technologies. One example, the Biosafety Clearing-House, was established by the 2000 Cartagena Protocol on Biosafety and serves as a publicly accessible repository of information on living modified organisms (LMOs) and on the genetic elements associated with those organisms.[341] The Clearing-House's objectives are to share information about LMO use and risk analyses, assist parties in making decisions about LMOs, provide evidence of treaty compliance, and foster international trade.[342]

One advantage of registries is that their establishment does not require coordinated global action. For example, ClinicalTrials.gov is a registry

---

338. *Id.*

339. *Id.*

340. *AI Alliance Launches as an International Community of Leading Technology Developers, Researchers, and Adopters Collaborating Together to Advance Open, Safe, Responsible AI*, IBM (Dec. 5, 2023), https://newsroom.ibm.com/AI-Alliance-Launches-as-an-International-Community-of-Leading-Technology-Developers,-Researchers,-and-Adopters-Collaborating-Together-to-Advance-Open,-Safe,-Responsible-AI [https://perma.cc/9UCG-SMHX].

341. *What Is the Biosafety Clearing-House (BCH)?*, BIOSAFETY CLEARING-HOUSE (Nov. 23, 2021), https://bch.cbd.int/en/kb/tags/about/What-is-the-Biosafety-Clearing-House-BCH-/619c553658029700017ff43b [https://perma.cc/4QY6-H8LD]; Cartagena Protocol on Biosafety to the Convention on Biological Diversity art. 20, Jan. 29, 2000, 2226 U.N.T.S. 208 [hereinafter Biosafety Protocol].

342. Tomme Rosanne Young, *Use of the Biosafety Clearing-House in Practice*, *in* LEGAL ASPECTS OF IMPLEMENTING THE CARTAGENA PROTOCOL ON BIOSAFETY 137–38 (Marie-Claire Cordonier et al. eds., 2013); *see also Human Genome Editing (HGE) Registry*, WORLD HEALTH ORG., https://www.who.int/groups/expert-advisory-committee-on-developing-global-standards-for-governance-and-oversight-of-human-genome-editing/registry [https://perma.cc/RVR2-D39X].

maintained by the U.S. National Library of Medicine that contains approximately 454,000 clinical studies from over 200 countries.[343] The registry allows researchers and patients from all over the world to identify relevant studies and research needs.[344] Over time, various organizations, including the World Medical Association and the International Committee of Medical Journal Editors, have adopted policies requiring registration in ClinicalTrials.gov or an equivalent registry.[345]

Registries could play an important role in promoting AI transparency, with different registries focusing on specific uses or concerns. A handful of cities are already using AI registries to inform residents about their use of AI systems.[346] China has instituted a semi-public, mandatory registry for algorithms involving recommendations, synthetic content generation, and generative AI.[347] Pending AI regulation in the European Union would require registration of high-risk AI systems in a public database.[348] Pennsylvania legislators have proposed a registry for businesses operating AI systems in the state,[349] and scientists have established a registry for AI in biomedical research to improve the quality and reproducibility of biomedical AIs.[350]

---

343. CLINICALTRIALS.GOV, https://clinicaltrials.gov [https://perma.cc/J3DG-AN7K]. The registry contains information about medical studies on human volunteers, including information about study protocols and outcomes.

344. *About ClinicalTrials.gov*, CLINICALTRIALS.GOV, https://beta.clinicaltrials.gov/about-site/about-ctg [https://perma.cc/N3PM-DBCF] (June 7, 2024).

345. *Id.*; *Clinical Trial Reporting Requirements*, CLINICALTRIALS.GOV, https://classic.clinicaltrials.gov/ct2/manage-recs/background#RegLawPolicies [https://perma.cc/689Y-GFYH] (June 7, 2024).

346. MEERI HAATAJA ET AL., PUBLIC AI REGISTERS: REALISING AI TRANSPARENCY AND CIVIC PARTICIPATION IN GOVERNMENT USE OF AI 3 (2020), https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf [https://perma.cc/6LZW-CY2K]; *AI Reviews & Algorithm Register*, CITY OF SAN JOSE, https://www.sanjoseca.gov/your-government/departments-offices/information-technology/digital-privacy/ai-reviews-algorithm-register [https://perma.cc/E47C-8U2Y].

347. Matt Sheehan, China's AI Regulations and How They Get Made 13 (July 2023) (working paper), https://carnegie-production-assets.s3.amazonaws.com/static/files/202307-Sheehan_Chinese%20AI%20gov-1.pdf [https://perma.cc/Q8SU-M9CD] (explaining that developers must submit information on how algorithms are trained and deployed and complete a security self-assessment report).

348. Michael Veale & Frederik Z. Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, 4 COMPUT. L. REV. INT'L 97, 111–12 (2021).

349. H.R. 49, 2023–2024 Leg., Reg. Sess. (Pa. 2023).

350. Julian Matschinske et al., *The AIMe Registry for Artificial Intelligence in Biomedical Research*, 18 NATURE METHODS 1128 (2021); *The AIMe Registry for Artificial Intelligence in Biomedical Research*, AIME REGISTRY, https://aime-registry.org [https://perma.cc/X23Q-LG4F].

In the context of AI safety, registries could be useful if they include AI developers, infrastructure providers, and large players.[351] A similar reporting mechanism for whistleblowers could also allow the reporting of suspected unethical or unsafe AI research or activities.[352] Such registries, if developed domestically, could serve as building blocks for international registries.[353]

On the other side, some aspects of AI developments should not be widely shared. Broad sharing of technological know-how would accelerate development, and for the many reasons we have outlined, this may be unsafe without rigorous safety and regulatory mechanisms. Note that registries do not have to be publicly open, and could confine disclosures to a regulatory body, rather than the public. The International Atomic Energy Agency ("IAEA") offers one example of an international organization that accesses and analyzes sensitive information while avoiding broader disclosure.[354]

2. Harmonization

Harmonizing regulatory requirements to reduce differences between regulatory regimes is a common objective of international governance. AI is the subject of intense international competition, and countries may fear that domestic regulation of AI development or deployment could put them at a strategic disadvantage.[355] Harmonization of AI regulation would counter incentives for countries to participate in a regulatory race to the bottom and for actors to relocate to jurisdictions with weaker regulations.[356] Harmonization would also facilitate the consideration of transboundary

---

351. UNESCO, MISSING LINKS IN AI GOVERNANCE 17–18 (Benjamin Prud'homme et al. eds., 2023).

352. *Cf.* World Health Organization [WHO], *Human Genome Editing: Recommendations*, at 14 (2021), https://iris.who.int/bitstream/handle/10665/342486/9789240030381-eng.pdf [https://perma.cc/8727-S77P] (recommending creation of "mechanism for confidential reporting of concerns about possibly illegal, unregistered, unethical and unsafe human genome editing research and other activities").

353. *Id.* at 18.

354. Allison Carnegie & Austin Carson, *The Disclosure Dilemma: Nuclear Intelligence and International Organizations*, 63 AM. J. POL. SCI. 269, 270, 274–78 (2019).

355. James S. Denford et al., *Weird AI: Understanding What Nations Include in Their Artificial Intelligence Plans*, BROOKINGS INST. (Apr. 25, 2023), https://www.brookings.edu/blog/techtank/2023/04/25/weird-ai-understanding-what-nations-include-in-their-artificial-intelligence-plans [https://perma.cc/WE52-XCES]; Rishi Iyengar, *The Global Race to Regulate AI*, FOREIGN POL'Y (May 5, 2023), https://foreignpolicy.com/2023/05/05/eu-ai-act-us-china-regulation-artificial-intelligence-chatgpt [https://perma.cc/PC36-QYT9].

356. Gary E. Marchant & Brad Allenby, *Soft Law: New Tools for Governing Emerging Technologies*, 73 BULL. ATOMIC SCIENTISTS 108, 109 (2017).

effects, reduce the potential for trade disputes, and ease regulatory burdens on multinational companies.[357]

Tools for promoting legal harmonization include registries and model standards. We already noted the Biosafety Clearing-House, which also collects information on national laws and regulations regarding the use and handling of LMOs, as well as decisions, risk assessments, and environmental reviews of such organisms.[358] The sharing of such information not only facilitates regulatory compliance but also enables countries to draw on others' efforts in developing their own regulatory systems and making regulatory decisions.[359]

Model regulatory standards can also promote harmonization. The World Health Organization, whose mission includes the establishment of international standards for pharmaceutical products, convenes expert committees to develop standards on good manufacturing practices, vaccines and biological products, and other subjects.[360] These standards have been adopted by countries and by the International Conference for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, which itself promulgates model standards for domestic adoption.[361]

As discussed below, various entities have developed a handful of technical standards for AI.[362] While yet to be fully implemented, these standards could play an important role in harmonization as jurisdictions grapple with how to regulate AI.

3. Technology Assessment

Assessments of emerging technologies can promote public engagement, identify risks, and analyze development trajectories and effects.[363] Policymakers and stakeholders can use the results of such assessments to manage risks and reshape the technologies themselves.[364] Performed

---

357. *Id.*

358. Biosafety Protocol, *supra* note 341, at 267.

359. Young, *supra* note 342, at 137–38.

360. VICTORIA WEISFELD & TRACY A. LUSTIG, INTERNATIONAL REGULATORY HARMONIZATION AMID GLOBALIZATION OF DRUG DEVELOPMENT: WORKSHOP SUMMARY 53 (2013).

361. *Id.*

362. *See infra* Section IV.C.4.

363. Albert C. Lin, *The Missing Pieces of Geoengineering Research Governance*, 100 MINN. L. REV. 2509, 2556–60 (2016).

364. *See* Albert C. Lin, *Technology Assessment 2.0: Revamping Our Approach to Emerging Technologies*, 76 BROOK. L. REV. 1309, 1349–50, 1353 (2011).

internationally or with international support, technology assessments can also offer additional support for regulatory harmonization.

Assessments by the Organisation for Economic Cooperation and Development ("OECD") have played a significant role in the international oversight of genetically modified organisms ("GMOs"). The OECD regularly prepares safety assessments of GMOs in the environment and foods derived from genetically modified crops.[365] The assessments do not obligate member countries to adopt a specific regulatory standard or any standard at all. Rather, these consensus documents aim to ensure that information used by member and non-member countries for GMO regulation is as similar as possible.[366] Establishing a common information base promotes more efficient risk assessment, harmonizes regulatory oversight, and reduces barriers to trade.[367] Although domestic regulation of GMOs exhibits substantial variation, the OECD assessments are widely read by regulators and industry and have been incorporated into the standard-setting work of international institutions.[368]

The experience with OECD assessments of GMOs suggests that assessments may be necessary but not sufficient to prompt regulatory harmonization—or even regulation—of emerging technologies. Consistent with this insight, Gary Marcus and Anka Reuel have proposed an "International Agency for AI" ("IAAI") that would include assessment as one of its core functions.[369] The IAAI's overarching mission would be to develop governance and technical solutions to promote safe AI technologies with the support of governments, business, nonprofits, and society at large.[370] To this end, the IAAI could collaboratively address problematic uses of AI, "convene experts and develop tools to tackle the spread of misinformation," and

---

365. *See Biosafety—BioTrack*, Org. for Econ. Coop. & Dev., https://www.oecd.org/chemicalsafety/biotrack [https://perma.cc/275D-K2LV].

366. *An Introduction to the Biosafety Consensus Documents of OECD's Working Group for Harmonisation in Biotechnology*, Organisation for Economic Co-operation and Development [OECD] 5, 8–9, ENV/JM/MONO(2005)5 (Feb. 22, 2005).

367. *Id.*

368. Helmut Gaugitsch, *The Impact of the OECD on the Development of National/International Risk/Safety Assessment Frameworks*, 5 ENV'T BIOSAFETY RES. 219, 221–22 (2006); Katharine Gostek, *Genetically Modified Organisms: How the United States' and the European Union's Regulations Affect the Economy*, 24 MICH. ST. INT'L L. REV. 761, 762, 782–84 (2016).

369. Gary Marcus & Anka Reuel, *The World Needs an International Agency for Artificial Intelligence, Say Two AI Experts*, ECONOMIST (Apr. 18, 2023), https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts; *see also* Bibek Debroy & Aditya Sinha, *Regulating Artificial Intelligence*, MERO TRIB. (Aug. 23, 2023), https://merotribune.com/2023/08/23/regulating-artificial-intelligence/ [https://perma.cc/7C9J-66D6].

370. Marcus & Reuel, *supra* note 369.

generate "swift and thoughtful guidance" from experts and researchers on responding to troubling developments.[371] Along these lines, the United Nations' High-Level Advisory Body on Artificial Intelligence has been tasked with "building a global scientific consensus on risks and challenges, helping harness AI for the Sustainable Development Goals, and strengthening international cooperation on AI governance."[372]

4. Soft Law

Soft law, as distinguished from enforceable hard law, refers to nonbinding standards.[373] Soft law includes principles, guidelines, codes of conduct, resolutions, certification and auditing requirements, and private standards developed by a wide range of institutions or governing bodies.[374] Soft law can be developed relatively quickly and is potentially applicable on an international scale.[375] It can also be an important step toward the formation of hard law, as international consensus builds around a soft law norm.[376] However, soft law itself lacks direct enforceability and accountability.[377] Indeed, because compliance is voluntary, soft law may suffer from a lack of participation by the bad actors whose compliance is most needed.[378] Nonetheless, indirect means can encourage or even mandate adherence to soft

---

371. *Id.*

372. Press Release, United Nations, UN Secretary-General Launches AI Advisory Body on Risks, Opportunities, and International Governance of Artificial Intelligence (Oct. 25, 2023), https://www.un.org/sites/un2.un.org/files/231025_press-release-aiab.pdf [https://perma.cc/2RKY-PS2G].

373. DANIEL BODANSKY, THE ART AND CRAFT OF INTERNATIONAL ENVIRONMENTAL LAW 14, 99 (2010); Marchant & Allenby, *supra* note 356, at 112; DAVID HUNTER ET AL., INTERNATIONAL ENVIRONMENTAL LAW & POLICY 339 (6th ed. 2022).

374. BODANSKY, *supra* note 373, at 14; Marchant & Allenby, *supra* note 356, at 112; Gary E. Marchant & Carlos I. Gutierrez, *Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence*, 24 MINN. J.L. SCI. & TECH. 375, 385 (2023); *see also* Rory Van Loo, *The Missing Regulatory State: Monitoring Businesses in an Age of Surveillance*, 72 VAND. L. REV. 1563 (2019) ("Dialogue would further allow government monitors to better comprehend complex algorithms. Regulatory monitors do not simply examine in silence, but as part of a dialectic process").

375. Marchant & Allenby, *supra* note 356, at 113.

376. *See* HUNTER ET AL., *supra* note 373, at 339.

377. GARY MARCHANT, "SOFT LAW" GOVERNANCE OF ARTIFICIAL INTELLIGENCE 15 (2019), https://escholarship.org/content/qt0jq252ks/qt0jq252ks.pdf?t=po1uh8 [https://perma.cc/ZRP5-EP2U].

378. *Id.* at 4.

law. Such indirect tools include certification programs, government procurement policies, and insurance contract provisions.[379]

A leading example of soft law is the Helsinki Guidelines, which set out ethical principles for medical research regarding human subjects. Adopted in 1964 by the World Medical Association, the Helsinki Guidelines have come to serve as "a central guide to research practice" and a foundation for other, more detailed ethical standards governing medical research.[380] Although the guidelines themselves are not legally binding, they are enforced indirectly through domestic laws that incorporate the guidelines and through journal publishers' demands that published research comply with the guidelines.[381]

Acknowledging the need for international oversight of AI, the U.N. Secretary-General has created a high-level advisory body to prepare initiatives on AI.[382] Although the form these initiatives might take is unclear, they will likely involve soft law. Indeed, soft law for AI has grown rapidly in recent years, even as measuring its actual implementation has proven difficult.[383]

Many soft law initiatives for AI have taken the form of principles proposed or developed by intergovernmental organizations, professional associations, and private entities.[384] The OECD, for example, has published five general "principles for responsible stewardship of trustworthy AI," accompanied by recommendations for national policies and international cooperation.[385] Another set of principles, the UNESCO Recommendation on the Ethics of Artificial Intelligence, calls for avoidance of unwanted harms, protection of privacy, and transparency and explainability in the deployment of AI.[386]

---

379. Marchant & Gutierrez, *supra* note 374, at 403–24.

380. Robert V. Carlson et al., *The Revision of the Declaration of Helsinki: Past, Present and Future*, 57 BRIT. J. CLINICAL PHARMACOLOGY 695, 704–05 (2004).

381. Delon Human & Sev S. Fluss, The World Medical Association's Declaration of Helsinki: Historical and Contemporary Perspectives 2–3 (Jan. 17, 2001) (unpublished manuscript), https://www.overgangsalderen.dk/wordpress/wp-content/uploads/2020/04/Declaration-of-Helsinki-Fifth-draft_historical_contemporary_perspectives-24-07-2001.pdf [https://perma.cc/Q62B-289D].

382. U.N. Advisory Body on A.I., Interim Report: Governing AI for Humanity (2023), https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf [https://perma.cc/H6TF-6NBF].

383. Marchant & Gutierrez, *supra* note 374, at 393, 424.

384. MARCHANT, *supra* note 377, at 5–10; Marchant & Gutierrez, *supra* note 374, at 393; *see also, e.g.*, *IBM's Principles for Trust and Transparency*, IBM, https://www.ibm.com/artificial-intelligence/ethics [https://perma.cc/3K2G-PXZX].

385. *Recommendation of the Council on Artificial Intelligence*, Organisation for Economic Co-operation and Development [OECD] 7–8, OECD/LEGAL/0449 (2022).

386. *Recommendation on the Ethics of Artificial Intelligence*, United Nations Educational, Scientific and Cultural Organization [UNESCO] 20–22, SHS/BIO/PI/2021/1 (2022).

These guidelines, which have been adopted by all 193 UNESCO member states, have been especially influential in developing countries.[387]

Soft law AI initiatives are not limited to the public sector.[388] The Partnership on AI, started by key industry players but now comprising academic, civil society, and media organizations as well,[389] has identified six "pillars"—"sets of issues where [the Partnership] sees some of the greatest risks and opportunities for AI"—and eight "tenets," such as "seek[ing] to ensure that AI technologies benefit and empower as many people as possible."[390]

As critics have noted, these principles tend to be general and difficult to operationalize.[391] However, other forms of soft law can provide more specific direction. Technical standards are process, design, or manufacturing specifications that—if well-designed and widely accepted—promote consistency and safety.[392] Technical standards typically reflect a consensus developed from expert consultations but often arise though closed processes that lack public input and democratic legitimacy.[393] A handful of technical standards for AI have been issued by the International Organization for Standardization ("ISO"), Institute of Electrical and Electronics Engineers ("IEEE"), and other entities.[394] The ISO, a nongovernmental organization composed of representatives of national standards bodies,[395] has issued several draft or final AI standards in partnership with the International Electrotechnical Committee, including standards for AI management systems (ISO 42001), AI governance (ISO 38507), and AI risk management (ISO

387. Melissa Hiekkila, *Our Quick Guide to the 6 Ways We Can Regulate AI*, MIT TECH. REV. (May 22, 2023), https://www.technologyreview.com/2023/05/22/1073482/our-quick-guide-to-the-6-ways-we-can-regulate-ai [https://perma.cc/X23W-GUZ7]; *Ethics of Artificial Intelligence*, UNESCO, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [https://perma.cc/Q42Y-842N].

388. Veale et al., *supra* note 334, at 5.

389. MARCHANT, *supra* note 377, at 7.

390. *About Us*, PARTNERSHIP ON AI, https://partnershiponai.org/about [https://perma.cc/9LAR-U6BB].

391. UNESCO, *supra* note 351, at 16.

392. Walter G. Johnson & Diana M. Bowman, *A Survey of Instruments and Institutions Available for the Global Governance of Artificial Intelligence*, 40 IEEE TECH. & SOC'Y MAG. 68, 71 (2021).

393. *Id.*; Veale et al., *supra* note 334, at 10.

394. Johnson & Bowman, *supra* note 392, at 71.

395. *What We Do*, ISO, https://www.iso.org/what-we-do.html [https://perma.cc/9EWM-PK8Z].

23894).[396] The IEEE has issued draft or final standards on subjects such as the transparency of autonomous systems, algorithmic bias, and addressing ethical concerns during system design.[397] The U.S. National Institute of Standards and Technology, a public entity, has also issued a voluntary framework for AI risk management.[398] In addition, the G7 has released a code of conduct for organizations developing advanced AI systems.[399] These various standards are increasingly serving as a starting point for efforts to develop domestic regulation.[400]

5. Hard Law

Treaties, conventions, and similar instruments constitute hard law—binding obligations of the states that enter into such agreements.[401] A hard law approach to AI could initially establish procedural requirements that are easy to meet, such as disclosing how systems are monitored, their operators registered, and their training runs audited—and later incorporate substantive

---

396. Hadrien Pouget, *What Will the Role of Standards Be in AI Governance?*, ADA LOVELACE INST. (Apr. 5, 2023), https://www.adalovelaceinstitute.org/blog/role-of-standards-in-ai-governance [https://perma.cc/7PK3-DH6B]; *ISO/IEC 23894:2023(en)*, INT'L ORG. FOR STANDARDIZATION, https://www.iso.org/obp/ui/en/#iso:std:iso-iec:23894:ed-1:v1:en [https://perma.cc/LEG6-8KSR]; Sam De Silva & Barbara Zapisetskaya, *Managing AI: What Businesses Should Know About the Proposed ISO Standard*, CMS LAW-NOW (Apr. 14, 2023), https://cms-lawnow.com/en/ealerts/2023/04/managing-ai-what-businesses-should-know-about-the-proposed-iso-standard [https://perma.cc/X3H5-DPL2].

397. *See IEEE Introduces Free Access to AI Ethics and Governance Standards*, LIBR. LEARNING SPACE: ACCESS, https://librarylearningspace.com/ieee-introduces-free-access-to-ai-ethics-and-governance-standards [https://perma.cc/7G2H-9RHP]; *see also* Alan F.T. Winfield et al., *IEEE P7001: A Proposed Standard on Transparency*, FRONTIERS ROBOTICS & AI (July 26, 2021), https://www.frontiersin.org/articles/10.3389/frobt.2021.665729/full [https://perma.cc/A6HM-M6F2]; JOSEP SOLER GARRIDO ET AL., AI WATCH: ARTIFICIAL INTELLIGENCE STANDARDISATION LANDSCAPE UPDATE 4–5 (2023).

398. NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COM., NIST AI 100-1, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [https://perma.cc/YRG7-RVU2].

399. G7 2023 HIROSHIMA SUMMIT, HIROSHIMA PROCESS INTERNATIONAL CODE OF CONDUCT FOR ORGANIZATIONS DEVELOPING ADVANCED AI SYSTEMS (2023), https://www.mofa.go.jp/files/100573473.pdf [https://perma.cc/VG8V-5L9C]; *G7 Leaders' Statement on the Hiroshima AI Process*, WHITE HOUSE (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/ [https://perma.cc/32XL-QHCM].

400. Pouget, *supra* note 396.

401. HUNTER ET AL., *supra* note 373, at 285.

standards as appropriate.[402] Treaties typically do not apply to non-state entities, however, and monitoring and enforcement may be ineffective.[403] Furthermore, negotiating and ratifying treaties take significant time and resources, and modifying treaties in response to new developments or information is likewise difficult.[404] These complexities pose a challenge to treaty governance in rapidly developing fields such as AI.[405]

Domestic regulation can have transnational impacts and offers a likely starting point for developing international AI regulation.[406] While legislatures have enacted dozens of laws that mention AI, many of these laws focus on specific applications of AI, and not all seek to regulate it.[407] Nonetheless, growing momentum to regulate AI nationally, as well as stakeholder and public support for AI regulation, suggest the feasibility of global AI oversight.[408] At the national level, overall approaches to AI regulation fall into three basic categories: applying existing law, devising new regulations that categorize AI applications by risk, and establishing requirements for testing and approval before use.[409]

Looking to position itself "as an AI superpower," the United Kingdom is following the first approach.[410] The United Kingdom directs regulators to apply a principles-based AI framework, in combination with existing law, on a context-specific basis.[411] Rather than regulating AI as a general matter, regulators are to consider specific uses of AI and incorporate principles such as safety, fairness, and transparency into the application of existing rules to AI.[412] While AI-specific legislation might be adopted if necessary, the

---

402. *How to Worry Wisely About Artificial Intelligence*, ECONOMIST (Apr. 20, 2023), https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence; Bill Whyman, *AI Regulation Is Coming—What Is the Likely Outcome?*, CTR. FOR STRATEGIC & INT'L STUD. (Oct. 10, 2023), https://www.csis.org/blogs/strategic-technologies-blog/ai-regulation-coming-what-likely-outcome [https://perma.cc/X9DN-HNUQ].

403. BODANSKY, *supra* note 373, at 15–16, 157.

404. Marchant & Allenby, *supra* note 356, at 110.

405. MARCHANT, *supra* note 377, at 3.

406. Veale et al., *supra* note 334, at 12.

407. Shana Lynch, *2023 State of AI in 14 Charts*, STAN. UNIV. HUMAN-CENTERED A.I. (Apr. 3, 2023), https://hai.stanford.edu/news/2023-state-ai-14-charts [https://perma.cc/B8YC-XX8U].

408. David Marchese, *How Do We Ensure an A.I. Future that Allows for Human Thriving?*, N.Y. TIMES (May 2, 2023), https://www.nytimes.com/interactive/2023/05/02/magazine/ai-gary-marcus.html (reporting comments by NYU professor Gary Marcus regarding bipartisan and global support for international regulation of AI).

409. *How to Worry Wisely About Artificial Intelligence*, *supra* note 402.

410. DEPARTMENT FOR SCIENCE, INNOVATION AND TECHNOLOGY, A PRO-INNOVATION APPROACH TO AI REGULATION, 2023, Cm. 815, at 2 (UK).

411. *Id.* at 5–6, 19, 25, 35.

412. *Id.* at 26–27.

approach relies heavily on existing law, as complemented by soft law in the form of technical standards and assurance techniques.[413] This approach falls short of what is needed in several regards—most notably its avoidance of general technology regulation and its blindness to societal-level risks. Still, it marks political will and interest in regulation of some kind.

The European Union, by contrast, is in the process of adopting a tiered, risk-based approach.[414] The EU Artificial Intelligence Act "categorizes applications of AI into four levels of risk: unacceptable risk, high risk, limited risk[,] and minimal or no risk."[415] Applications involving unacceptable risk, such as AI systems using manipulative or deceptive techniques to distort behavior and untargeted scraping of facial images to create facial recognition databases, are prohibited.[416] High-risk applications, which include use of AI systems to influence elections and systems that may cause significant potential harm to health, safety, fundamental rights, and the environment, are subject to manufacturer assessment of impacts on fundamental rights as well as other requirements.[417] Limited risk applications, including deepfakes and chatbots, are subject to minimal transparency obligations.[418] For minimal or no risk applications, member states are encouraged to apply voluntary codes

---

413. *Id.* at 29, 56.

414. Kim Mackrael, *Sweeping Regulation of AI Advances in European Union Deal*, WALL ST. J. (Dec. 8, 2023), https://www.wsj.com/tech/ai/regulation-of-ai-advances-in-european-union-deal-09d18355 (explaining that political deal reached on AI regulation in December 2023 still requires final approval from parliamentarians and representatives); Jess Weatherbed, *Why the AI Act Was So Hard to Pass*, VERGE (Dec. 13, 2023), https://www.theverge.com/2023/12/13/23999849/eu-ai-act-artificial-intelligence-regulations-complicated-delays [https://perma.cc/RC5L-66RB] (noting that E.U. agreement on AI regulation is based on principles and that approved text of AI act is still being crafted).

415. Ryan Browne, *Europe Takes Aim at ChatGPT with What Might Soon Be the West's First A.I. Law. Here's What It Means*, CNBC (May 15, 2023), https://www.cnbc.com/2023/05/15/eu-ai-act-europe-takes-aim-at-chatgpt-with-landmark-regulation.html [https://perma.cc/NB5N-27NK].

416. European Parliament Press Release, Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI (Dec. 9, 2023), https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai [https://perma.cc/84LW-S3SL]; Council of the European Union Press Release 986/23, Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for AI in the World (Dec. 9, 2023), https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/pdf [https://perma.cc/RJS8-3QY3]. The legislation allows use of biometric identification systems for law enforcement purposes in targeted searches involving specified serious crimes. European Parliament Press Release, *supra*.

417. European Parliament Press Release, *supra* note 416; Veale & Borgesius, *supra* note 348, at 102–06.

418. Veale & Borgesius, *supra* note 348, at 106.

of conduct.[419] In addition, general-purpose AI systems are subject to transparency obligations, as well as risk assessment and mitigation and other requirements if they involve high impacts and systemic risk.[420] The European Union's approach nonetheless fails to address misalignment concerns and to capture several high-risk categories. It does not apply to AI systems used for military or defense purposes, including autonomous weapons systems.[421] It also does little to address concerns about systems that can autonomously and recursively self-improve.[422] Yet, we should also acknowledge that this early action illustrates a strong political will and interest in transnational regulation.

China has taken a somewhat more restrictive approach with respect to targeted AI applications. Building on its registration requirements for specified AI algorithms, China issued an interim regulation for generative AI in July 2023.[423] Under this interim approach, providers of AI services to the public for generating text, images, audio, video, or other content "bear responsibility as the producers of online information content."[424] Providers must "[e]mploy effective measures to increase the quality of training data, and increase the truth, accuracy, objectivity, and diversity" of such data.[425] Furthermore, providers of "generative AI services with public opinion properties or the capacity for social mobilization" must carry out and submit "security assessments" to regulators before making such services publicly available.[426] The regulation also includes privacy, transparency, and accountability requirements,[427] as well as a requirement that generated content "[u]phold the Socialist Core Values."[428] Notably, the regulation

---

419. *Id.* at 98.

420. European Parliament Press Release, *supra* note 416.

421. Council of the European Union Press Release 986/23, *supra* note 416.

422. *Id.*

423. Sheehan, *supra* note 347, at 14.

424. *Interim Measures for the Management of Generative Artificial Intelligence Services*, CHINA L. TRANSLATE art. 9 (July 13, 2023) [hereinafter *Interim Measures*], https://www.chinalawtranslate.com/en/generative-ai-interim/ [https://perma.cc/K8LY-U96C].

425. *Id.* art. 7.

426. *Id.* art. 17; *see also* Josh Ye & Urvi Manoj Dugar, *China Lets Baidu, Others Launch ChatGPT-Like Bots to Public, Tech Shares Jump*, REUTERS (Aug. 31, 2023), https://www.reuters.com/technology/baidu-among-first-win-china-approval-ai-models-bloomberg-news-2023-08-30/ [https://perma.cc/MH99-CRPP].

427. *Interim Measures*, *supra* note 424, arts. 4, 7, 10, 11, 15, 19; Matt O'Shaughnessy, *What a Chinese Regulation Proposal Reveals About AI and Democratic Values*, CARNEGIE ENDOWMENT FOR INT'L PEACE (May 16, 2023), https://carnegieendowment.org/2023/05/16/what-chinese-regulation-proposal-reveals-about-ai-and-democratic-values-pub-89766 [https://perma.cc/8GD4-QVFD].

428. *Interim Measures*, *supra* note 424, art. 4(1).

applies only to the private sector, not to governmental use of AI.[429] As a result, some observers worry China's development and use of AI for national security, surveillance, and military purposes will proceed unabated.[430]

Aspects from each of these approaches might be incorporated into global AI standards. Depending on the desired functions of governance, international AI governance may take distinct forms in different contexts. For some AI applications, coordination and harmonization of standards will take priority. In such instances, the International Civil Aviation Organization ("ICAO") might serve as an appropriate model for international governance.[431] This U.N. agency, charged with fostering the development of international air transport, establishes standards and recommended practices for international air navigation.[432]

In other contexts, managing the risks posed by AI will be of foremost concern, requiring a more vigorous approach. In this vein, various stakeholders have suggested that the IAEA might serve as a model for AI regulation.[433] "Focus[ed] on reducing existential risk," an IAEA-like entity could "inspect systems, require audits, test for compliance with safety standards, [and] place restrictions on degrees of deployment and levels of security."[434] Alternatively, a global AI regulator might have a more limited sphere of responsibility, such as focusing on the use of autonomous weapons.[435]

---

429. O'Shaughnessy, *supra* note 427. The regulations apply only to the provision of generative AI services to the public, and not to research and development or internal use within companies. *See* Mark MacCarthy, *The US and Its Allies Should Engage with China on AI Law and Policy*, BROOKINGS INST. (Oct. 19, 2023), https://www.brookings.edu/articles/the-us-and-its-allies-should-engage-with-china-on-ai-law-and-policy/ [https://perma.cc/5JQ3-EHWP].

430. *See* Sigal Samuel, *The Case for Slowing Down AI*, VOX (Mar. 20, 2023, 7:58 AM EDT), https://www.vox.com/the-highlight/23621198/artificial-intelligence-chatgpt-openai-existential-risk-china-ai-safety-technology [https://perma.cc/VL85-G42W].

431. *See* Marcus & Reuel, *supra* note 369 (describing the ICAO as a "softer kind of model, with less focus on enforcement").

432. *About ICAO*, INT'L CIV. AVIATION ORG., https://www.icao.int/about-icao/Pages/default.aspx [https://perma.cc/Y4U6-NXQS].

433. Altman et al., *supra* note 240; Press Release, Secretary-General, *supra* note 382 (noting that the IAEA "is a model that could be very interesting" because it "is a very solid, knowledge-based institution" that "has some regulatory functions"); Marcus & Reuel, *supra* note 369 (identifying IAEA as a possible precedent for global cooperation).

434. Altman et al., *supra* note 240.

435. *See* Kai-Fu Lee, *The Third Revolution in Warfare*, ATLANTIC (Sept. 11, 2021), https://www.theatlantic.com/technology/archive/2021/09/i-weapons-are-third-revolution-warfare/620013 [https://perma.cc/VS3P-KKDQ] (discussing regulation of, or ban on, autonomous weapons, as potential responses to danger of autonomous weapons arms race);

While the IAEA can provide a useful precedent for international AI regulation, distinctions between nuclear proliferation and AI suggest that AI governance will be more complex. The IAEA regulates state actors, its inspection and monitoring activities assume the ability to detect physical nuclear material, and its role evolved over decades in response to revealed gaps in oversight.[436] By contrast, any AI oversight system will have to account for AI development and use by both private actors and states across a wide range of sectors.[437] AI efforts will likely be more difficult to detect because they lack the substantial physical footprint of nuclear weapons.[438] While GPU server farms do leave a footprint, distributed training paradigms may enable sophisticated actors to evade detection. Furthermore, AI is developing rapidly, leaving less time for the gradual evolution of a governance structure.[439]

International governance of AI need not require an international regulator, however. An international treaty could spell out binding obligations to be implemented by individual states, without oversight from an international monitor. For example, the Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, adopted by the Council of Europe in May 2024, obligates states to ensure that AI systems incorporate individual privacy protections, transparency and auditability requirements, and safety and security requirements.[440] The treaty opens for signature on September 5, 2024, and could be signed by not only the forty-six member states of the Council, but also observer states—including the United States, Mexico, and Japan.[441]

UNESCO, *supra* note 351, at 333, 337–38 (urging adoption of a binding treaty to prohibit antipersonnel autonomous weapons and regulating other uses of autonomous weapons).

436. Ian J. Stewart, *Why the IAEA Model May Not Be Best for Regulating Artificial Intelligence*, BULL. ATOMIC SCIENTISTS (June 9, 2023), https://thebulletin.org/2023/06/why-the-iaea-model-may-not-be-best-for-regulating-artificial-intelligence/ [https://perma.cc/9W4R-7RAL].

437. *Id.*; Huw Roberts et al., *Global AI Governance: Barriers and Pathways Forward*, 100 INT'L AFFS. 1275, 1282 (May 7, 2024), https://academic.oup.com/ia/article/100/3/1275/7641064 [https://perma.cc/UDU5-2DYZ].

438. Stewart, *supra* note 436.

439. *See id.*

440. *Council of Europe Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*, COUNCIL OF EUR. (May 17, 2024), https://rm.coe.int/1680afae3c [https://perma.cc/E5GG-Q4EC]; *see* Hannah van Kolfschooten & Carmel Shachar, *The Council of Europe's AI Convention (2023–2024): Promises and Pitfalls for Health Protection*, 138 HEALTH POL'Y 104935 (2023).

441. Hiekkilä, *supra* note 387.

Ongoing efforts to develop oversight and accountability mechanisms for AI, whether in the form of registries, principles, technical standards, or domestic law, reflect the accretion of an AI governance network. These various mechanisms are laying the foundation for international governance of AI. Strengthening connections between key players in governance can facilitate information-sharing, coordination, and norm-building.[442] While establishing binding and meaningful international governance of AI may prove challenging, precedents in other areas indicate that such governance is achievable and normatively desirable.

V. CONCLUSION

This Article lays out the case for the broad, systemic regulation of AI. The dangers of AI systems extend to present and future harms. They range from fraud and misinformation to property damage and human lives. They threaten communities and they may involve national or transnational threats. Our principal argument is that all these risks matter. To mitigate these risks and allow society to reap the benefits of this new technology, comprehensive government regulation will be necessary.

The present AI moment already exposes a sliver of the full dangers of AI systems. Their broad deployment threatens bias and discrimination on a new scale, the erosion of social trust, and uncomfortable threats to privacy when algorithms can infer our intimate secrets. As AI systems gain new capabilities, they may have transformative effects on labor markets with resulting impacts on wealth and inequality. Their military applications can be used to make violence efficient and accurate to an unprecedented degree. And their power could engender new modes of surveillance and totalitarianism.

These threat profiles largely stem from misuse by AI system engineers. But these systems can also cause massive social harms due to their own misalignment. We have detailed the alignment problem and noted that we should expect that even systems pursuing benign goals will impose considerable social risks. Solving the alignment problem, however, turns out to be more complex than most realize. It is a problem that we currently do not know how to solve.

We see both benefits and risks in the future development and deployment of AI systems. We have demonstrated that, even on a conventional cost-benefit basis, the case for regulation is strong. Recognizing uncertainty does not alter that; rather, reasonable precaution demands that future development

---

442. Roberts et al., *supra* note 437, at 13–14.

be even more tightly regulated. To that end, we have provided a set of regulatory recommendations, based on both a domestic and an international strategy. We explored a set of seven principles that domestic regulation should follow. We also explored international precedents and noted the important role of a combination of transparency and secrecy. We also demonstrated that international cooperation is indeed plausible and highlighted a variety of examples to that effect.

Ultimately, every honest assessment must start and end with epistemic humility. We simply do not know many things, and we do not always know the things that we do not know. But if there is a deep uncertainty over whether a plane is safe or not, it is best not to board it.[443] AI systems promise power. It is the hardest thing to resist. Market participants would like to assure us that they will use it responsibly and will not deploy systems that are unsafe. They would like to see, if anything, regulation that focuses on bits and parcels, and only on specific applications. We believe that there is a role for robust, systemic regulation, and that an informed policy conversation about the risks and upsides of AI will point the way toward the optimal regulatory approach. We hope to have started that conversation here.

443. NASSIM NICHOLAS TALEB, ANTIFRAGILE: THINGS THAT GAIN FROM DISORDER 160 (2012).