ALGORITHMIC BLACK SWANS

Noam Kolt[*]

101 WASH. U. L. REV. (forthcoming)

From biased lending algorithms to chatbots that spew violent hate speech, AI systems already pose many risks to society. While policymakers have a responsibility to tackle pressing issues of algorithmic fairness, privacy, and accountability, they also have a responsibility to consider broader, longer-term risks from AI technologies. In public health, climate science, and financial markets, anticipating and addressing societal-scale risks is crucial. As the COVID-19 pandemic demonstrates, overlooking catastrophic tail events — or "black swans" — is costly. The prospect of automated systems manipulating our information environment, distorting societal values, and destabilizing political institutions is increasingly palpable. At present, it appears unlikely that market forces will address this class of risks. Organizations building AI systems do not bear the costs of diffuse societal harms and have limited incentive to install adequate safeguards. Meanwhile, regulatory proposals such as the White House AI Bill of Rights and the European Union AI Act primarily target the immediate risks from AI, rather than broader, longer-term risks. To fill this governance gap, this Article offers a roadmap for "algorithmic preparedness" — a set of five forward-looking principles to guide the development of regulations that confront the prospect of algorithmic black swans and mitigate the harms they pose to society.

[*] Vanier Scholar and Doctoral Candidate, University of Toronto Faculty of Law; Graduate Affiliate, Schwartz Reisman Institute for Technology and Society; Research Affiliate, Legal Priorities Project. For helpful comments and suggestions, I thank Anthony Niblett, John Bliss, Brian Frye, Isaac Gazendam, Gillian Hadfield, Riley Harris, Matthijs Maas, Hadrian Pouget, Jonas Schuett, Michal Shur-Ofry, Malcolm Thorburn, Risto Uuk, Suzanne Van Arsdale, José Jaime Villalobos, Albert Yoon, and discussants at the 2023 Consumer Law Scholars Conference, 2023 Privacy Law Scholars Conference, Emory University School of Law Workshop on Vulnerability Theory and Digital Intimacy, Cornell Law School Inter-University Graduate Conference, Siena-Tel-Aviv-Toronto Law and Economics Workshop, and UC Berkeley Center for Human-Compatible AI.

TABLE OF CONTENTS

INTRODUCTION

On November 30, 2022, OpenAI released ChatGPT, an AI chatbot that can engage in human-like dialogue and perform a diverse range of complex tasks.[1] The chatbot, developed by one of the world's leading AI research labs, can write essays and emails, generate and debug computer code, and explain concepts in physics and philosophy.[2] Within two months of its release, ChatGPT amassed one hundred million users.[3] The underlying technology, which has countless applications, is anticipated to become as widespread and influential as search engines and smartphones.[4]

But transformative technologies like ChatGPT have a dark underbelly.[5]

---

[1] John Schulman et al., *ChatGPT: Optimizing Language Models for Dialogue*, OPENAI (Nov. 30, 2022), https://openai.com/blog/chatgpt/.

[2] *Id.*

[3] Krystal Hu, *ChatGPT Sets Record for Fastest-Growing User Base*, REUTERS (Feb. 2, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[4] *See* Ethan Mollick, *ChatGPT Is a Tipping Point for AI*, HARV. BUS. REV. (Dec. 14, 2022), https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai; Cade Metz & Nico Grant, *A New Chat Bot Is a 'Code Red' for Google's Search Business*, N.Y. TIMES (Dec. 21, 2022), https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html. Notably (and perhaps infamously), Microsoft's Bing uses OpenAI's GPT-4 language model. *See* Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html; Yusuf Mehdi, *Confirmed: The New Bing Runs on OpenAI's GPT-4*, MICROSOFT BING BLOGS (Mar. 14, 2023), https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4.

[5] For discussion of the risks from language model technologies, which power ChatGPT, see Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, PROC. 2021 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610 (2021); Laura Weidinger et al., *Ethical and Social Risks of Harm from Language Models*, ARXIV at 9–35 (Dec. 8, 2021), https://arxiv.org/abs/2112.04359; Laura Weidinger et al., *Taxonomy of Risks Posed by Language Models*, PROC. 2022 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 214 (2022); Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV at 128–59 (Aug. 16, 2021), https://arxiv.org/abs/2108.07258; OpenAI, *GPT-4 Technical Report*, ARXIV at 44–65, (Mar. 15, 2023), https://arxiv.org/abs/2303.08774; Irene Solaiman et al., *Evaluating the Social Impact of Generative AI Systems in Systems and Society*, ARXIV (June 12, 2023), https://arxiv.org/abs/2306.05949. For broader discussion of the risks associated with AI technologies, see Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 10–18 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION chs. 2, 4 (2015); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–93 (2016); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY chs. 3–10 (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1027–34 (2017); Ryan Calo, *Artificial Intelligence Policy: A*

The models powering chatbots are trained on vast quantities of text scraped from the internet, as well as feedback from human crowdworkers.[6] As a result of patterns in these data, the chatbot can amplify harmful biases, including by generating violent hate speech and producing discriminatory hiring algorithms.[7] In addition, tools like ChatGPT routinely provide users with inaccurate and misleading responses, which could pollute or systematically manipulate our information environment.[8]

Some of these harms have already materialized. Stack Overflow, a popular online forum for computer programmers, explained that while answers produced by ChatGPT appear reliable, they are often erroneous.[9] Given the ease of generating enormous quantities of these answers, and the platform's inability to vet the quality of answers on a case-by-case basis, Stack Overflow judged the risk unacceptably high and decided to ban all answers generated by ChatGPT.[10]

Primer and Roadmap, 51 U.C. DAVIS L. REV. 399, 411–27 (2017); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 26–29 (2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR ch. 4 (2018); MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN chs. 2–3, 5 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2227–62 (2019); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 138–48 (2019); FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI chs. 4–6 (2020); KATE CRAWFORD, THE ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE (2021); Daren Acemoglu, *Harms of AI* (NBER Working Paper No. 29247, Sept. 2021), https://www.nber.org/papers/w29247.

[6] For discussion of the psychological harm suffered by these workers, see Billy Perrigo, *OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, TIME (Jan. 18, 2023), https://time.com/6247678/openai-chatgpt-kenya-workers/. For discussion of issues concerning the prevailing method for training AI models with human feedback, known as RLHF, see Stephen Casper et al., *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*, ARXIV (Sept. 11, 2023), https://arxiv.org/abs/2307.15217.

[7] *See* Sam Biddle, *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*, THE INTERCEPT (Dec. 8, 2022), https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

[8] *See* Melissa Heikkilä, *How to Spot AI-Generated Text*, MIT TECH. REV. (Dec. 19, 2022) ("In an already polarized, politically fraught online world, these AI tools could further distort the information we consume. If they are rolled out into the real world in real products, the consequences could be devastating."). *See also* Avid Ovadya, *What's Worse Than Fake News? The Distortion of Reality Itself*, 35 NEW PERSP. Q. 43 (2018) (describing the "catastrophic failure of the marketplace of ideas" as an "infopocalypse").

[9] Site Moderators, *Use of ChatGPT Generated Text for Content on Stack Overflow Is Temporarily Banned*, STACK OVERFLOW (Dec. 5, 2022), https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned.

[10] *Id.* ("Overall, because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and

This incident is a canary in a coalmine.[11] Society is likely to face a host of new challenges as programmers, journalists, business leaders, and politicians increasingly use AI systems that appear competent but are fundamentally untrustworthy.[12] Due to the unprecedented scale at which systems like ChatGPT operate, the resulting harms will be widespread. For example, automated systems that interact with millions of people could subtly manipulate societal values and even influence the outcomes of political processes.[13]

Computer scientists identify two reasons why these risks are likely to intensify. First, more competent AI systems are not necessarily more trustworthy.[14] For example, a recent study found that more powerful models

---

to users who are asking or looking for correct answers.") *See also* Editors, *Tools Such As ChatGPT Threaten Transparent Science*, 613 NATURE 612 (2023) (imposing a prohibition on crediting language models as authors of scientific papers); H. Holden Thorp, *ChatGPT Is Fun, But Not an Author*, 379 SCIENCE 313 (2023); Abeba Birhane, Atoosa Kasirzadeh, David Leslie & Sandra Wachter, *Science in the Age of Large Language Models*, 5 NATURE REV. PHYS. 277 (2023).

[11] *See* James Vincent, *AI-generated Answers Temporarily Banned on Coding Q&A Site Stack Overflow*, VERGE (Dec. 5, 2022), https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers ("The worry is that this pattern could be repeated on other platforms, with a flood of AI content drowning out the voices of real users with plausible but incorrect data. Exactly how this could play out in different domains around the web, though, would depend on the exact nature of the platform and its moderation capabilities."); Gary Marcus, *AI's Jurassic Park Moment*, THE ROAD TO AI WE CAN TRUST (Dec. 10, 2022), https://garymarcus.substack.com/p/ais-jurassic-park-moment (explaining that "systems like [ChatGPT] pose a real and imminent threat to the fabric of society" because they are "inherently unreliable, frequently making errors of both reasoning and fact," "can easily be automated to generate misinformation at unprecedented scale", and "cost almost nothing to operate").

[12] *See* Wendy Pollack, *ChatGPT Holds Promise and Peril*, WASH. POST (Dec. 18, 2022), https://www.washingtonpost.com/business/energy/chatgpt-holds-promise-and-peril/2022/12/17/e74a2aa0-7e13-11ed-bb97-f47d47466b9a_story.html. Even the CEO of OpenAI acknowledged the chatbot's significant shortcomings. *See* Sam Altman (@sama), TWITTER (Dec. 11, 2022), https://twitter.com/sama/status/1601731295792414720?lang=en ("ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. it's [sic.] a mistake to be relying on it for anything important right now. it's a preview of progress; we have lots of work to do on robustness and truthfulness.").

[13] *See infra* Part IV.C (exploring the impact of AI on social and political institutions).

[14] *See infra* Part I.B (illustrating that many AI systems are unsafe by default). *See also* Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety*, ARXIV (June 21, 2016), https://arxiv.org/ abs/1606.06565; STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019); BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2020); Richard Ngo, *AGI Safety from First Principles* (Sept. 2020) (manuscript on file with author); Dan Hendrycks, Nicholas Carlini, John Schulman & Jacob Steinhardt, *Unsolved Problems in ML Safety*, ARXIV (Sept. 28, 2021), https://arxiv.org/abs/2109.13916; Richard Ngo, Lawrence Chan & Sören Mindermann, *The*

have a greater tendency to produce responses that reinforce a user's own preferences and, thereby, bolster ideological echo chambers.[15] Second, despite these concerns, users are increasingly willing to deploy AI systems in high-stakes settings.[16] While potentially beneficial, these applications can backfire. For instance, tools developed to automate drug discovery can be repurposed to design chemical weapons.[17] Meanwhile, malfunctioning AI systems that control critical infrastructure or administer access to essential services could have disastrous consequences.[18]

Societal risks on this scale are known as "black swans."[19] Nassim Taleb, who popularized the term in his study of financial crises, describes black swans as unexpected extreme-impact events, that is, highly consequential risks that are difficult to predict ex ante but easy to explain in hindsight.[20] Examples in the context of financial markets, public health, and geopolitics include the 2008 Great Recession, the COVID-19 global pandemic, and Russia's 2022 invasion of Ukraine, respectively.[21] This Article explores

---

Alignment Problem from a Deep Learning Perspective, ARXIV (Dec. 16, 2022), https://arxiv.org/abs/2209.00626; Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, PROC. 2023 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 651 (2023); Dan Hendrycks, Mantas Mazeika & Thomas Woodside, *An Overview of Catastrophic AI Risks*, ARXIV (Sept. 11, 2023), https://arxiv.org/abs/2306.12001. For related discussion in the context of language models, see Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik & Geoffrey Irving, *Alignment of Language Agents*, ARXIV (Mar. 26, 2021), https://arxiv.org/abs/2103.14659; Bommasani et al., *supra* note 5, at 113–16; Weidinger et al., *Ethical and Social Risks*, *supra* note 5, at 10; Weidinger et al., *Taxonomy*, *supra* note 5; Solaiman et al., *supra* note 5.

[15] Ethan Perez et al., *Discovering Language Model Behaviors with Model-Written Evaluations*, FINDINGS 2023 CONF. ASS'N COMPUT. LINGUISTICS 133387, 13392–93 (2023) (describing this phenomenon as language model "sycophancy").

[16] Hendrycks, Carlini, Schulman & Steinhardt, *supra* note 14, at 1–2.

[17] *See* Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins, *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTELL. 189, 189 (2022) ("In less than 6 hours after starting on our in-house server … the AI designed not only VX, but also many other known chemical warfare agents").

[18] *See, e.g.*, Bommasani et al., *supra* note 5, at 115–16.

[19] *See* NASSIM NICHOLAS TALEB, THE BLACK SWAN: THE IMPACT OF THE HIGHLY IMPROBABLE xvii (2007) (the term is inspired by European ornithologists' "discovery" of black swans in Australia, having previously thought that all swans were white). For further discussion of the term's origins, see Sanat Pai Raikar, *Black Swan Event*, BRITANNICA (updated Aug. 18, 2023), https://www.britannica.com/topic/black-swan-event.

[20] *Id*. at xvii–xviii (describing the three features of a black swan: "First, it is an *outlier*, as it lies outside the realm of regular expectations, because nothing in the past can convincingly point to its possibility. Second, it carries an extreme impact. Third, in spite of its outlier status, human nature makes us concoct explanations for its occurrence *after* the fact, making it explainable and predictable. I stop and summarize the triplet: rarity, extreme impact, and retrospective (though not prospective) predictability.").

[21] *But see* Bernard Avishai, *The Pandemic Isn't a Black Swan but a Portent of a More Fragile Global Systems*, NEW YORKER (Apr. 21, 2020) (discussing Taleb's opposition to

another class of high-impact societal risk, arising from the widespread adoption of unsafe AI technology: *algorithmic black swans*.

These high-impact risks from AI are hard to predict because technological progress is hard to predict. The field of AI is characterized by sudden, often unexpected developments.[22] Quantitative changes can lead to qualitatively different capabilities. For example, building larger models and using larger datasets have enabled AI systems to translate between languages, produce photorealistic images, and answer bar exam questions.[23] Just as these capabilities can emerge without warning, ethical and social harms can occur suddenly and in surprising contexts. For instance, text generation tools can produce toxic outputs in response to seemingly benign inputs.[24] AI systems can also pursue goals that are different from, and even antithetical to, societal interests, giving rise to the so-called "alignment problem."[25]

Misaligned AI systems have already caused grave harm in criminal justice, healthcare, and other sensitive settings.[26] For example, recidivism prediction tools have exhibited racial biases,[27] medical chatbots have promoted self-harm,[28] and automated trading algorithms have caused

---

describing the COVID-19 pandemic as a black swan).

[22] *See infra* Part I.A (discussing the emergent capabilities of AI systems).

[23] *See* Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo Arredondo, *GPT-4 Passes the Bar Exam*, (Working Paper, Apr. 5, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233; Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan & Daniel Schwarcz, *ChatGPT Goes to Law School* (Working Paper, May 19, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335905. *But see* Eric Martinez, *Re-Evaluating GPT-4's Bar Exam Performance* (Working Paper, Sept. 26, 2023), https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=4441311.

[24] Deep Ganguli et al., *Predictability and Surprise in Large Generative Models*, PROC. 2022 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1747, 1750 (2022).

[25] *See* RUSSELL, *supra* note 14; CHRISTIAN, *supra* note 14; Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACH. 411 (2020); Kenton, Everitt, Weidinger, Gabriel, Mikulik & Irving, *supra* note 14; Ngo, Chan & Mindermann, *supra* note 14; Atoosa Kasirzadeh & Iason Gabriel, *In Conversation with Artificial Intelligence: Aligning Language Models with Human Values*, 36 PHIL. & TECH 27 (2023); Iason Gabriel & Vafa Ghazavi, *The Challenge of Value Alignment: From Fairer Algorithms to AI Safety*, *in* THE OXFORD HANDBOOK OF DIGITAL ETHICS (Carissa Véliz ed., forthcoming); Anton Korinek & Avital Balwit, *Aligned with Whom? Direct and Social Goals for AI Systems* (NBER Working Paper No. 30017, May 2022), https://www.nber.org/papers/w30017.

[26] *See* PASQUALE, *supra* note 5; Barocas & Selbst, *supra* note 5; O'NEIL, *supra* note 5; NOBLE, *supra* note 5; EUBANKS, *supra* note 5; Kleinberg, Ludwig, Mullainathan & Sunstein, *supra* note 5; CRAWFORD, *supra* note 5; Acemoglu, *supra* note 5.

[27] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (uncovering the COMPAS algorithm's anti-Black bias).

[28] *See* Anne-Laure Rousseau, Clément Baudelaire & Kevin Riera, *Doctor GPT-3: Hype or Reality?*, NABLA (Oct. 27, 2020), https://www.nabla.com/blog/gpt-3/ (revealing that the GPT-3 language model recommended that a hypothetical patient commit suicide); Lauren

financial turmoil.[29] While significant resources have been allocated to tackling these well-documented harms from AI, far fewer resources have been allocated to addressing larger-scale societal harms, such as risks to critical infrastructure and democratic institutions.[30]

Although we might hope that as AI systems play increasingly important roles in society, computer scientists and software developers will take steps to mitigate the risk of algorithmic black swans, current market dynamics suggest otherwise.[31] The AI industry overwhelmingly prioritizes improving the capabilities of systems, not their safety or social impact. The prevailing culture is one of unrelenting progress, not caution. This need not be the case. By analogy, civil engineers are not tasked with building "safe bridges." Rather, safety is an inherent part of bridgebuilding.[32] In contrast, AI developers generally under-invest in safety, building systems without sufficient guardrails and relegating safety to an afterthought.[33]

This market failure stems from two main factors. First, leading AI labs face significant pressure to outpace their competitors in building systems that exhibit state-of-the-art performance irrespective of the ethical and societal consequences.[34] Second, organizations building AI systems are unlikely to bear the social cost of harms caused by the technologies they create.[35]

Walker, *Belgian Man Dies by Suicide Following Exchanges with Chatbot*, BRUSSELS TIMES (Mar. 28, 2023), https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

[29] *See* Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi & Tugkan Tuzun, *The Flash Crash: High-Frequency Trading in an Electronic Market*, 72 J. FIN. 967 (2017) (describing an incident in 2010 in which automated trading systems triggered a trillion-dollar stock market crash).

[30] *See infra* Part IV (illustrating that current regulatory proposals neglect some of the most consequential societal risks posed by AI). For further discussion of the priorities of the AI ethics and governance communities, see Stephen Cave & Seán S. ÓhÉigeartaigh, *Bridging Near- and Long-term Concerns about AI*, 1 NATURE MACH. INTELL. 5 (2019); Carina Prunkl & Jess Whittlestone, *Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society*, PROC. 2020 AAAI/ACM CONF. AI, ETHICS & SOC'Y 138 (2020).

[31] *See infra* Part II.A (discussing the priorities of AI researchers and developers).

[32] *See* Hendrycks, Carlini, Schulman & Steinhardt, *supra* note 14, at 13 ("Safety is not auxiliary in most current widely deployed technology. … Their safety is insisted upon—even assumed—and incorporating safety features is imbued in the design process. The [machine learning] community should similarly create a culture of safety and elevate its standards so that [machine learning] systems can be deployed in safety-critical situations.").

[33] *See, e.g.*, MUSTAFA SULEYMAN, THE COMING WAVE: TECHNOLOGY, POWER, AND THE TWENTY-FIRST CENTURY'S GREATEST DILEMMA ch. 14 (2023) ("Safety features should not be afterthoughts but inherent design properties of all these new technologies, the ground state of everything that comes next.").

[34] *See infra* Part II.B (discussing the impact of competition on the safety practices of AI researchers and developers).

[35] *See infra* Part II.C (examining the externalities generated by unsafe AI systems).

Clearly, the case for regulatory intervention is strong. The challenge is to design interventions that effectively tackle the most concerning societal risks.

Regulators in the United States and Europe have responded to this challenge by proposing a host of new laws and policies for regulating AI.[36] These range from "soft law" initiatives in the United States, such as the White House's Blueprint for an AI Bill of Rights[37] and the National Institute of Standards and Technology's AI Risk Management Framework,[38] to "hard law" proposals in the European Union, including the EU AI Act[39] and the EU AI Liability Directive.[40] While these instruments address some of the immediate risks posed by AI, they contain notable gaps with respect to broader and longer-term risks from the technology.

Three gaps stand out. The first gap concerns *general purpose AI systems*, that is, AI systems such as ChatGPT that can perform a diverse range of tasks across different domains. Because these systems serve as "foundation models" that underpin many downstream applications,[41] including in high-stakes settings, a failure to operate ethically or safely could be catastrophic.[42] The second gap concerns *proliferation and misuse*. As AI technologies diffuse widely and rapidly, they can easily be adapted for malicious purposes, such as orchestrating large-scale cyberattacks and perpetrating financial fraud.[43] The third gap concerns *systemic risk*. In addition to harming large numbers of people, AI systems can cause severe damage to social and political institutions. For instance, the proliferation of biased text generation

[36] *See infra* Part III (surveying recent U.S. and EU proposals for regulating AI). *See also* Margot Kaminski, *Regulating the Risks from AI*, 103 B.U. L. REV. 101 (forthcoming 2023).

[37] White House Office of Sci. & Tech. Policy, *Blueprint for an AI Bill of Rights: Making Automated Systems Work* (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [hereinafter White House AI Bill of Rights].

[38] National Institute of Standards and Technology, *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework [hereinafter NIST AI RMF].

[39] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM (2021) 206 final (Apr. 21, 2021) [hereinafter EU AI Act].

[40] European Commission, Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) COM (2022) 496 final (Sept. 28, 2022) [hereinafter EU Liability Directive].

[41] *See* Bommasani et al., *supra* note 5, at 3 (coining the term "foundation model"); *Huge "Foundation Models" Are Turbo-charging AI Progress*, ECONOMIST (Jul. 11, 2022), https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress.

[42] *See infra* Part IV.A (discussing the societal risks posed by general purpose AI systems).

[43] *See infra* Part IV.B (examining the proliferation and misuse of AI technologies).

tools could exacerbate political polarization.[44] Regulatory proposals in the United States and Europe largely fail to address these algorithmic black swans.

What can, and should, regulators do differently? How can they more effectively prevent AI systems from causing large-scale societal harm? Drawing on insights from public health, financial regulation, and climate policy, this Article offers a roadmap for *algorithmic preparedness*—a set of five forward-looking principles to guide the development of regulations that address the risk of algorithmic black swans.[45]

The first principle ("anticipation") concerns the *goals* of regulating AI technologies: AI governance should aim to, among other things, anticipate and mitigate large-scale societal harm from AI systems—a goal that is neglected by current regulatory proposals. The subsequent principles ("diversification," "scalability," and "experimentation") concern the *means* for achieving that goal: AI governance should adopt a portfolio approach comprised of diverse, uncorrelated, and highly scalable regulatory strategies, while continually exploring and evaluating new regulatory strategies.[46] The final principle ("recalibrating risk") concerns balancing the benefits and costs of AI regulation. It suggests that cost-benefit analysis of AI governance interventions should place greater weight on worst-case outcomes.

Implementing these principles will involve complex questions of regulatory design. How can regulators obtain up-to-date and accurate information about the capabilities and impact of advanced AI systems? Which interventions are most likely to incentivize AI developers to increase their investment in safety? What institutional frameworks can provide reliable feedback on the effectiveness of different governance strategies? This Article tackles these questions with humility. The longer-term societal challenges from AI, including black swans, entail great uncertainty. Clarifying the goals of AI governance and developing a framework for more targeted and robust intervention is a good place to start.

The remainder of this Article is organized as follows. Part I explores the emergent capabilities and risks of current AI systems. Despite being unsafe by default, these systems are nonetheless being deployed in increasingly complex and sensitive settings. Part II examines the market dynamics affecting the development and use of AI, finding that companies are not

---

[44] *See infra* Part IV.C (illustrating that the widespread adoption of unsafe AI systems could cause substantial harm to social and political institutions).

[45] *See infra* Part V (describing the five principles of algorithmic preparedness and illustrating how policymakers can implement them in practice).

[46] Some companies have adopted this approach to AI safety research. *See, e.g.*, Anthropic, *Core Views on AI Safety: When, Why, What, and How*, ANTHROPIC (Mar. 8, 2023), https://www.anthropic.com/index/core-views-on-ai-safety.

incentivized to mitigate the risk of algorithmic black swans. Part III provides an overview of U.S. and EU proposals for regulating AI. Part IV illustrates that these proposals contain notable gaps and fail to address large-scale societal risks. Part V outlines five principles for improving algorithmic preparedness and mitigating the risk of algorithmic black swans.

I. EMERGENT RISKS

A. More is Different

The past decade has seen remarkable progress in AI technology, introducing machines that can perform ever more complex and diverse tasks. The era's defining moment arrived in 2012, when University of Toronto computer scientist Geoffrey Hinton and his graduate students built a neural network that achieved unprecedented performance in recognizing images.[47] Hinton subsequently received the Turing Award, the highest honor in computer science.[48] In addition to establishing neural networks as the bedrock architecture for AI systems,[49] Hinton and his team revealed the importance of scale: building *larger* models trained on *larger* datasets with *larger* computational resources is the key to unlocking the capabilities of neural networks.

---

[47] *See* Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, PROC. 25TH CONF. NEURAL INFO. PROCESSING SYS. 1097 (2012). In 2022, the paper was unanimously selected as the NeurIPS Test of Time paper. *See* Sahra Ghalebikesabi, *Announcing the NeurIPS 2022 Awards* (Nov. 21, 2022), https://blog.neurips.cc/2022/11/21/announcing-the-neurips-2022-awards/. *See also* CADE METZ, GENIUS MAKERS: THE MAVERICKS WHO BROUGHT AI TO GOOGLE, FACEBOOK, AND THE WORLD ch. 5 (2021) (recounting the history of Hinton's academic group).

[48] *See Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award*, ASSOC. COMPUT. MACH. (2018), https://awards.acm.org/about/2018-turing (jointly awarding the prize to Yoshua Bengio, Geoffrey Hinton, and Yann LeCun). Notably, two of the recipients—Bengio and Hinton—endorsed a public statement concerning societal-scale risks from AI. *See Statement on AI Risk*, CENTER FOR AI SAFETY (May 30, 2023), https://www.safe.ai/statement-on-ai-risk. Bengio also signed a letter calling on AI companies to pause training of models larger than GPT-4. *See Pause Giant AI Experiments: An Open Letter*, FUTURE OF LIFE INST. (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Hinton left his role at Google to more openly discuss the risks from AI. *See* Cade Metz, *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead*, N.Y. TIMES (May 1, 2023), https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

[49] *See* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436 (2015).

One of Hinton's graduate students, Ilya Sutskever, took this lesson to heart.[50] After revolutionizing the field of machine translation at Google,[51] Sutskever went on to co-found OpenAI, a company with the bold ambition of building artificial general intelligence, that is, creating AI systems that can perform a broad range of economically valuable activities.[52] In 2020, another revolution took place. OpenAI released GPT-3, a machine learning model capable of completing college-level exams, generating computer code, and producing fluent human-like prose.[53] What set GPT-3 apart from earlier, less capable models? The answer is *scale*.[54] The immense size of the model and its training data gave rise to new capabilities. Quantitative changes led to qualitatively different results. In short, the model showed that *more is different*.[55]

Notably, many of the newly discovered abilities of AI systems have come as a surprise.[56] Progress in AI is not linear, but erratic.[57] It is often difficult to predict how changes to a model's inputs will affect its performance. For example, merely extending the length of training enabled a model to proceed from utterly failing to answer logic and math questions to achieving near-perfect accuracy.[58] These sudden jumps in performance can be likened to

---

[50] METZ, *supra* note 57, at 145 (quoting Sutskever: "The real conclusion is that if you have a very large dataset and a very large neural network then success is guaranteed.").

[51] *See* Gideon Lewis-Kraus, *The Great A.I. Awakening*, N.Y. TIMES (Dec. 14, 2016) (describing the integration of Sutskever's research into Google Translate).

[52] *See OpenAI Charter*, OPENAI (Apr. 9, 2018), https://openai.com/charter/ ("OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI…").

[53] *See* Brown et al., *Language Models Are Few-Shot Learners*, PROC. 34TH CONF. NEURAL INFO. PROCESSING SYS. 1877 (2020) (introducing the GPT-3 language model).

[54] *See* Jason Wei et al., *Emergent Abilities of Large Language Models*, TRANSACTIONS MACH. LEARN. RES. at 2 (Aug. 2022) (discussing the scaling of several AI resources); Nestor Maslej et al., *The AI Index 2023 Annual Report*, STANFORD UNIVERSITY INSTITUTE FOR HUMAN-CENTERED AI at 56, 60 (Apr. 2023), https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (documenting increases in the size of AI models and computational resources).

[55] The phrase was popularized by Nobel prize-winning physicist Philip Anderson. *See* Philip Anderson, *More is Different*, 177 SCIENCE 393 (1972). *See also* Jacob Steinhardt, *More Is Different for AI*, BOUNDED REGRET (Jan. 4, 2022), https://bounded-regret.ghost.io/more-is-different-for-ai/.

[56] *See* Ganguli et al., *supra* note 24; Wei et al., *supra* note 54. *But see* Rylan Schaeffer, Brando Miranda & Sanmi Koyejo, *Are Emergent Abilities of Large Language Models a Mirage?*, ARXIV (May 22, 2023), https://arxiv.org/abs/2304.15004.

[57] Scaling laws, however, can sometimes predict the impact of scaling on performance. *See, e.g.*, Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ARXIV (Jan. 23, 2020), https://arxiv.org/abs/2001.08361; Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, ARXIV (Mar. 29, 2022), https://arxiv.org/abs/2203.15556.

[58] *See* Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin & Vedant Misra,

phase transitions in physical phenomena, such as water freezing or boiling when it reaches a certain temperature.[59] By analogy, the performance of an AI system can radically improve when a critical threshold is reached. A chess-playing agent, for example, underwent a phase transition at a certain point in its training, during which the model spontaneously learned the concepts of king safety, threats, and mobility.[60]

The challenge of discovering the emergent abilities of AI systems has led some researchers to suggest that today's AI systems might contain a "capabilities overhang."[61] That is, these systems may be far more capable than we assume. After all, the only capabilities we observe are those that we actively test and benchmark. Other capabilities can go undetected. The same is true for safety risks. For example, an AI agent trained to play the strategy game *Diplomacy* unexpectedly learned to deceive and manipulate its human opponents.[62] Meanwhile, OpenAI's GPT-4 model successfully recruited a human crowdworker to complete a CAPTCHA task designed to distinguish between humans and bots.[63] Protecting against unknown risks, arising from unknown capabilities, is notoriously difficult.[64]

Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, 1ST MATH. REASONING IN GENERAL AI WORKSHOP, INT'L CONF. LEARNING REPRESENTATIONS (2021).

[59] *See* Alexander Pan, Kush Bhatia & Jacob Steinhardt, *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models*, INT'L CONF. LEARNING REPRESENTATIONS at 2, 9 (2022). *See also* Wei et al., *supra* note 54, at 2 (describing phase transitions in AI systems as "a dramatic change in overall behavior that would not have been foreseen by examining smaller-scale systems").

[60] *See* Thomas McGrath et al., *Acquisition of Chess Knowledge in AlphaZero*, 119 PROC. NAT'L ACAD. SCI. e2206625119 at 6 (2022).

[61] *See* Jack Clark, *Import AI 310* (Nov. 28, 2022), https://jack-clark.net/2022/11/28/import-ai-310-alphazero-learned-chess-like-humans-learn-chess-capability-emergence-in-language-models-demoscene-ai/.

[62] Specifically, CICERO, an AI agent trained by Facebook's parent company Meta, agreed with another player not to attack a particular territory, but proceeded to "backstab" that player. *See* Zvi Mowshowitz, *On the Diplomacy AI*, DON'T WORRY ABOUT THE VASE (Nov. 28, 2022), https://thezvi.substack.com/p/on-the-diplomacy-ai. *See also* Meta Fundamental AI Research Diplomacy Team et al., *Human-level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning*, 378 SCIENCE 1067, app. 3–4 (2022) (discussing the risk of AI agents manipulating humans). For broader discussion of AI manipulation and deception, see Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, ARXIV (Aug. 28, 2023), https://arxiv.org/abs/2308.14752; Micah Carroll, Alan Chan, Henry Ashton & David Krueger, *Characterizing Manipulation from AI Systems*, ARXIV (Mar. 17, 2023), https://arxiv.org/abs/2303.09387;

[63] OpenAI, *GPT-4 Technical Report*, *supra* note 5, at 55–66; ARC Evals, *Update on ARC's Recent Eval Efforts* (Mar. 17, 2023), https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

[64] *See* Hendrycks, Carlini, Schulman & Steinhardt, *supra* note 14, at 7; Samuel R. Bowman, *The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP*

B. Unsafe by Default

It would be reassuring if improvements in the capabilities of AI systems were accompanied by equal improvements in their safety. This, however, is rarely the case. While the capabilities and safety of AI systems are sometimes correlated, in many instances they are not.[65] For example, systems that can produce better explanations of ideas in physics and philosophy can also produce more compelling misinformation.[66] There are also scenarios in which greater capabilities *decrease* safety.[67] For instance, more powerful language models (i.e., models that typically perform better on a wide range of language-related tasks) have been found to produce less truthful responses to certain questions, compared with weaker models.[68] In other words, a model that might be entrusted to perform more complex or sensitive tasks is more likely to mislead people.

Evidently, there can be a tradeoff between AI capabilities and AI safety.[69] As AI systems become more capable and are deployed in higher-stakes settings, this tradeoff could compound. Consider, for example, an AI system used to optimize energy usage in critical infrastructure, such as healthcare facilities, water treatment plants, or public transport systems. The potential upside—reducing carbon emissions on a massive scale—is tremendous.[70]

Systems Fail, PROC. 60TH ANN. MEETING ASS'N COMPUT. LINGUISTICS 7484, 7489 (2022) (arguing that understanding current AI capabilities is key to confronting the associated risks).

[65] This is related to the "orthogonality thesis," according to which the cognitive capabilities of an AI system and its goals vary independently of each other. *See*, *e.g.*, Stuart Armstrong, *General Purpose Intelligence: Arguing the Orthogonality Thesis*, 12 METAPHYSICS & ANALYSIS 68 (2013). *See also* Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 62–86 (2019) (discussing the inevitability of errors in software).

[66] *See infra* Part IV.C (discussing the risk of AI systems being used to generate and disseminate misinformation).

[67] This is sometimes described as "inverse scaling." *See* Ian R. McKenzie et al., *Inverse Scaling: When Bigger Isn't Better*, ARXIV (June 15, 2023), https://arxiv.org/abs/2306.09479.

[68] *See* Stephanie Lin, Jacob Hilton & Owain Evans, *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, PROC. 60TH ANN. MEETING ASS'N COMPUT. LINGUISTICS 3214, 3220 (2022). *See also* Perez et al., *Discovering Language Model Behaviors*, *supra* note 15, at 13393 (finding that larger models exhibit a greater tendency to behave "sycophantically," i.e., reinforce a user's existing preferences).

[69] *See* Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research*, ARXIV at 8–9 (Sept. 20, 2022), https://arxiv.org/abs/2206.05862. *See also* Jonas B. Sandbrink, Hamish Hobbs, Jacob L. Swett, Allan Dafoe & Anders Sandberg, *Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks* (Working Paper, Oct. 17, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670 (discussing differential technological development, which advocates a risk-reducing approach to developing AI and other technologies).

[70] *See*, *e.g.*, Richard Evans & Jim Gao, *DeepMind AI Reduces Google Data Centre*

But so are the potential downsides. The consequences of AI causing critical infrastructure to malfunction could be catastrophic.[71]

Researchers in the emerging field of AI safety, which focuses on mitigating such risks, suggest that we cannot assume AI systems are safe by default.[72] In fact, they argue the opposite. The issue, often described as the "alignment problem,"[73] is one of optimization: *how can we ensure that AI systems reliably optimize prosocial goals?* The problem can be divided into two parts. The first part concerns *specifying appropriate goals*.[74] Continuing with the example above, if the goal of an AI were to reduce energy usage in water treatment facilities (an ostensibly reasonable prosocial goal), it may

---

Cooling Bill by 40%, DEEPMIND (Jul. 20, 2016), https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40 (describing an AI system that dramatically reduced energy usage in commercial data centers).

[71] *See* Zachary Arnold & Helen Toner, *AI Accidents: An Emerging Threat*, GEORGETOWN CENTER FOR SECURITY & EMERGING TECHNOLOGY at 7–15 (Jul. 2021), https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/; Mia Hoffmann & Heather Frase, *Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework*, at 12–13, GEORGETOWN CENTER FOR SECURITY & EMERGING TECHNOLOGY (2023), https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

[72] *See generally* Amodei, Olah, Steinhardt, Christiano & Mané, *supra* note 14; RUSSELL, *supra* note 14; CHRISTIAN, *supra* note 14; Ngo, *supra* note 14; Hendrycks, Carlini, Schulman & Steinhardt, *supra* note 14; Ngo, Chan & Mindermann, *supra* note 14; Weidinger et al., *Ethical and Social Risks*, *supra* note 5; Weidinger et al., *Taxonomy*, *supra* note 5; Chan et al., *supra* note 5; Solaiman et al., *supra* note 5.

[73] *See* RUSSELL, *supra* note 14; CHRISTIAN, *supra* note 14; Gabriel, *supra* note 25; Ngo, Chan & Mindermann, *supra* note 14; Kasirzadeh & Gabriel, *supra* note 25; Gabriel & Ghazavi, *supra* note 25; Korinek & Balwit, *supra* note 25.

[74] *See* Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, DEEPMIND (Apr. 21, 2020), https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity ingenuity ("when rewarded for doing well on a homework assignment, a student might copy another student to get the right answers, rather than learning the material – and thus exploit a loophole in the task specification. This problem also arises in the design of artificial agents. For example, a reinforcement learning agent can find a shortcut to getting lots of reward without completing the task as intended by the human designer. … These problems are likely to become more challenging in the future, as AI systems become more capable at satisfying the task specification at the expense of the intended outcome.") For a canonical (and accessible) illustration of the phenomenon, see Jack Clark & Dario Amodei, *Faulty Reward Functions in the Wild*, OPENAI (Dec. 21, 2016), https://openai.com/blog/faulty-reward-functions/ (revealing that an agent trained to maximize the score in a video game engages in highly unexpected behavior). There is a burgeoning literature on goal misspecification and misgeneralization. *See* Lauro Langosco Di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau & David Krueger, *Goal Misgeneralization in Deep Reinforcement Learning*, 162 PROC. MACH. LEARNING. RES. 12004 (2022); Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato & Zac Kenton, *Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals*, ARXIV (Nov. 2, 2022), https://arxiv.org/abs/2210.01790; Pan et al., *The Effects of Reward Misspecification*, *supra* note 59.

altogether cease providing power to those facilities.[75] This is clearly an undesirable outcome. The second part of the problem concerns whether an AI system *in fact* optimizes the specified goal.[76] For example, a system that learns to optimize energy usage by selectively shutting down water treatment facilities in situations that go unnoticed by human operators is highly undesirable.

Aspects of the alignment problem are familiar to lawyers and social scientists. For lawyers, the challenge can be likened to *principal-agent problems*, which are pervasive in corporate governance, employment relationships, and contractual arrangements.[77] An agent (the AI system) is supposed to take actions that are in the best interests of a principal (the human designer or user, or a group of humans). Effectively incentivizing the agent and overseeing the agent's actions is very costly. For social scientists, the alignment problem recalls *Goodhart's law*: "when a measure becomes a target, it ceases to be a good measure."[78] In a familiar context, if shareholder value is measured solely by profit, a corporation may take socially noxious

---

[75] For discussion of the risks arising from automated control of energy facilities, see Amodei, Olah, Steinhardt, Christiano & Mané, *supra* note 14, at 16; Thomas Krendl Gilbert, Sarah Dean, Tom Zick & Nathan Lambert, *Choices, Risks, and Reward Reports Charting Public Policy for Reinforcement Learning Systems*, UC BERKELEY CENTER FOR LONG-TERM CYBERSECURITY at 35 (Feb. 2022), https://cltc.berkeley.edu/reward-reports/.

[76] *See* Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse & Scott Garrabrant, *Risks from Learned Optimization in Advanced Machine Learning Systems*, ARXIV (Dec. 1, 2021), https://arxiv.org/abs/1906.01820 (coining the term "mesa-optimization" to describe the problem). *See also* Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov & David Krueger, *Defining and Characterizing Reward Hacking*, 36TH CONF. NEURAL INFO. PROCESSING SYS. (2022); Michael K. Cohen, Marcus Hutter & Michael A. Osborne, *Advanced Artificial Agents Intervene in the Provision of Reward*, 43 AI MAG. 282 (2022); Leo Gao, John Schulman & Jacob Hilton, *Scaling Laws for Reward Model Overoptimization*, 202 PROC. MACH. LEARN. RES. 10835 (2023).

[77] *See* Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, PROC. 2019 AAAI /ACM CONF. AI, ETHICS & SOC'Y 417, 420–21 (2019) (applying insights from incomplete contracting theory to the problem of AI alignment). *See also* Sebastian Benthall & David Shekman, *Designing Fiduciary Artificial Intelligence*, ARXIV (Jul. 27, 2023), https://arxiv.org/abs/2308.02435 (analogizing the problem of AI alignment to fiduciary relationships). *But see* Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel, *Cooperative AI: Machines Must Learn to Find Common Ground* 593 NATURE 33 (2021) (illustrating that AI safety does not only concern the control of a single agent by a single person, but the interactions between multiple agents and multiple people).

[78] *See* Marilyn Strathern, *Improving Ratings: Audit in the British University System*, 5 EURO. REV. 305, 308 (1997), paraphrasing the idea introduced in Charles E. Goodhart, *Problems of Monetary Management: The U.K. Experience*, *in* PAPERS IN MONETARY ECONOMICS (1975). *See also* David Manheim & Scott Garrabrant, *Categorizing Variants of Goodhart's Law*, ARXIV (Feb. 24, 2019), https://arxiv.org/abs/1803.04585; Steven Kerr, *On the Folly of Rewarding A, While Hoping for B*, 18 ACAD. MGMT. J. 769 (1975).

actions provided those actions are profitable. Similarly, an AI agent whose performance is measured solely by a crude proxy may, in its efforts to maximize that proxy, cause tremendous societal harm.[79]

As AI systems perform higher-stakes activities, the alignment problem could become more acute. Consider, for example, current AI systems that augment the work of human programmers by automatically generating computer code.[80] A programmer typically uses these systems by entering some code of their own, following which the model "completes" the sequence by generating new code. Researchers have discovered that when provided with inputs that contain bugs or vulnerabilities, code generation systems are more likely to produce new code that also contains bugs and vulnerabilities.[81] This concerning phenomenon is, fundamentally, an alignment problem. Optimizing to produce new code that *mimics* the input code provided by fallible (human) programmers is not a desirable end-goal, but a dangerous proxy. Left unaddressed, this problem could grossly undermine computer security and cause vast economic damage.[82]

---

[79] *See* RUSSELL, *supra* note 14, at 140 ("One of the most common patterns involves omitting something from the objective that you do actually care about. In such cases … the AI system will often find an optimal solution that sets the thing you do care about, but forgot to mention, to an extreme value.") *See also* Stuart Russell, *Of Myths and Moonshine*, EDGE (Nov. 14, 2014), https://www.edge.org/conversation/the-myth-of-ai ("This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want."). *See also* Amodei, Olah, Steinhardt, Christiano & Mané, *supra* note 14, at 4 ("for an agent operating in a large, multifaceted environment, an objective function that focuses on only one aspect of the environment may implicitly express indifference over other aspects of the environment. An agent optimizing this objective function might thus engage in major disruptions of the broader environment if doing so provides even a tiny advantage for the task at hand.") These ideas are formalized in Simon Zhuang & Dylan Hadfield-Menell, *Consequences of Misaligned AI*, PROC. 34TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 15763 (2020).

[80] Examples of these systems include OpenAI's Codex and DeepMind's AlphaCode. *See* Mark Chen et al., *Evaluating Large Language Models Trained on Code*, ARXIV (July 14, 2021), https://arxiv.org/abs/2107.03374; Yujia Li et al., *Competition-level Code Generation with AlphaCode*, 378 SCIENCE 1092 (2022).

[81] *See* Chen et al., *supra* note 80, at 27. *See also* Neil Perry, Megha Srivastava, Deepak Kumar & Dan Boneh, *Do Users Write More Insecure Code with AI Assistants?*, ARXIV (Dec. 16, 2022), https://arxiv.org/abs/2211.03622; Erik Jones & Jacob Steinhardt, *Capturing Failures of Large Language Models via Human Cognitive Biases*, 36TH CONF. NEURAL INFO. PROCESSING SYS. at 1 (2022) ("Codex errs predictably based on how the input prompt is framed, adjusts outputs towards anchors, and is biased towards outputs that mimic frequent training examples.").

[82] *See*, *e.g.*, Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt & Ramesh Karri, *Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions*, PROC. 2022 IEEE SYMPOSIUM ON SECURITY & PRIVACY 754 (2022). For broader discussion of AI-related security risks, see Clark Barrett et al., *Identifying and Mitigating the Security Risks of Generative AI*, ARXIV (Aug. 28, 2023),

C. Black Swans

A sizeable fraction of the AI community is concerned about automated systems causing large-scale societal harm.[83] According to Stuart Russell, co-author of the most widely used textbook on AI,[84] safety is set to become a central priority of the field: "Just as nuclear fusion researchers consider the problem of containment of fusion reactions as one of the primary problems of their field, it seems inevitable that issues of control and safety will become central to AI as the field matures."[85]

Unpredictable high-impact risks should no longer surprise us. The first edition of Nassim Taleb's *Black Swan* immediately preceded the 2007–2008 financial crisis.[86] The cost of the COVID-19 pandemic and Russia's 2022 invasion of Ukraine—in human lives and economic value—has sensitized us to the potential magnitude of catastrophic tail risks. As political scientist Scott Sagan quipped: "things that have never happened before happen all the time."[87] Algorithmic black swans are no exception.

---

https://arxiv.org/abs/2308.14840.

[83] Recent surveys of AI researchers illustrate consistent concern about AI systems causing catastrophic harm. *See* Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang & Owain Evans, *When Will AI Exceed Human Performance? Evidence from AI Experts*, 62 J. AI RES. 729, 733 (2018); Baobao Zhang, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, Michael C. Horowitz, *Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers*, ARXIV at 7 (June 8, 2022), https://arxiv.org/abs/2206.04132; Zach Stein-Perlman, Benjamin Weinstein-Raun & Katja Grace, *2022 Expert Survey on Progress in AI*, AI IMPACTS (Aug. 3, 2022), https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/; Julian Michael et al., *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey*, ARXIV at 11 (Aug. 26, 2022), https://arxiv.org/abs/2208.12852. *See also supra* note 48. Several high-profile AI researchers, however, downplay the scale of risks posed by AI. *See*, *e.g.*, Caleb Garling, *Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines*, WIRED (May 2015), https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/; Yann LeCun, FACEBOOK (Feb. 23, 2016), https://www.facebook.com/ yann.lecun/posts/10153368458167143. Some legal scholars have expressed similar sentiments. *See*, *e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 431–33 (2017). *Compare* Kaminski, *supra* note 36, at 153 (suggesting that AI may present "wide-scale and catastrophic risks"). *See id.* at 69 ("we know that unlikely and potentially catastrophic events are more likely to happen with complex systems than with less complex technologies. We just can't measure or predict precisely what those events will be.").

[84] PETER NORVIG & STUART J. RUSSELL, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (4th ed. 2020).

[85] Stuart Russell, *The Long-term Future of AI*, UC BERKELEY DEPT. ELEC. ENG'G & COMP. SCI., https://people.eecs.berkeley.edu/~russell/research/future/.

[86] *See* TALEB, *supra* note 19.

[87] SCOTT D. SAGAN, THE LIMITS OF SAFETY: ORGANIZATIONS, ACCIDENTS, AND

Put simply, unpredictable high-impact technologies present unpredictable high-impact risks.[88] The risks presented by AI are diverse and growing. Automated systems that are entrusted with performing complex tasks in safety-critical settings could drastically amplify harmful biases and further entrench existing inequities.[89] Novel applications of AI, meanwhile, could give rise to new classes of risk. For example, autonomous systems used to optimize agricultural processes could cause large-scale crop failures or environmental degradation,[90] AI tools developed to accelerate drug discovery could be repurposed to design chemical weapons,[91] and AI systems that provide inaccurate or partial information to policymakers could cause dramatic societal harm.[92]

In each of these scenarios, black swan risks are not caused by a single technological artefact. Instead, the risks arise from the interaction of complex sociotechnical systems.[93] The decision to develop AI tools that optimize

---

NUCLEAR WEAPONS 12 (1993).

[88] *See* Amodei, Olah, Steinhardt, Christiano & Mané, *supra* note 14, at 2; Hendrycks, Carlini, Schulman & Steinhardt, *supra* note 14, at 3.

[89] *See* PASQUALE, *supra* note 5; Barocas & Selbst, *supra* note 5; O'NEIL, *supra* note 5; NOBLE, *supra* note 5; EUBANKS, *supra* note 5; Kleinberg, Ludwig, Mullainathan & Sunstein, *supra* note 5; CRAWFORD, *supra* note 5; Acemoglu, *supra* note 5.

[90] *See* Asaf Tzachor, Medha Devare, Brian King, Shahar Avin & Seán Ó hÉigeartaigh, *Responsible Artificial Intelligence in Agriculture Requires Systemic Understanding of Risks and Externalities*, 4 NATURE MACH. INTELL. 104, 105 (2022). For risks to critical water infrastructure, see Catherine E. Richards, Asaf Tzachor, Shahar Avin & Richard Fenner, *Rewards, Risks and Responsible Deployment of Artificial Intelligence in Water Systems*, 1 NATURE WATER 422 (2023).

[91] *See* Urbina, Lentzos, Invernizzi & Ekins, *supra* note 17. *See also* Justine Calma, *AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours*, VERGE (Mar. 12, 2022), https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx (quoting one of the paper's authors: "For me, the concern was just how easy it was to do. A lot of the things we used are out there for free. … If you have somebody who knows how to code in Python and has some machine learning capabilities, then in probably a good weekend of work, they could build something like this generative model driven by toxic datasets. So that was the thing that got us really thinking about putting this paper out there; it was such a low barrier of entry for this type of misuse.").

[92] *See* Gilbert, Dean, Zick & Lambert, *supra* note 75, at 48 ("Imagine the use of [a reinforcement learning] system that recommended the optimal policy for closing or re-opening private businesses to minimize COVID infections or related deaths. Here, both state actors and private vendors are functionally ignorant of system effects — the former unaware of technical choices underlying the tool's optimization, the latter lacking knowledge of the expert judgment needed to adjudicate use cases. [Reinforcement learning] may exacerbate this administrative gap, as private firms and startups develop AI systems that automate or appropriate functions of the state without anyone knowing until it is too late.").

[93] *See* Roel I. J. Dobbe, *System Safety and Artificial Intelligence*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, Baobao Zhang eds., forthcoming),

agricultural processes, accelerate drug discovery, or advise policymakers are value-laden choices affected by a combination of cultural factors, organizational structures, and regulatory environments.[94] Understanding the complex systems behind algorithmic black swans is perhaps the first step toward addressing this emerging risk.[95]

Sociotechnical complexity may also explain the relative neglect of algorithmic black swans, compared with other safety risks from AI, such as issues of fairness, transparency, and privacy.[96] Another explanation for researchers' reluctance to address large-scale risks from AI relates to black swans more generally. Human beings (including lawmakers) systematically overlook consequential tail events in many contexts.[97] Conventional risk

---

discussing NANCY G. LEVESON, ENGINEERING A SAFER WORLD: SYSTEMS THINKING APPLIED TO SAFETY (2016).

[94] *See* Laurin B. Weissinger, *AI, Complexity, and Regulation*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, Baobao Zhang eds., forthcoming); Matthijs M. Maas, *Aligning AI Regulation to Sociotechnical Change*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (Justin B. Bullock et al. eds., forthcoming).

[95] *See* Dobbe, *supra* note 93 (applying Leveson's systems engineering principles to AI safety and governance). For an application of systems thinking and complex systems theory to issues in technology law, see Kate Klonick, *Of Systems Thinking and Straw Me*n, 136 HARV. L. REV. F. 339, 340–41 (2023).

[96] There are entire computer science conferences dedicated to the latter issues, including the ACM Conference on Fairness, Accountability, and Transparency (FAccT). Increasingly, however, leading computer science conferences such as NeurIPS host workshops focused on tail risks from AI. *See, e.g.*, Workshop on Machine Learning Safety, NeurIPS 2022 https://nips.cc/virtual/2022/workshop/49986.

[97] *See* RICHARD A. POSNER, CATASTROPHE: RISK AND RESPONSE 8 (2004) ("In the case of law, neglect of the catastrophic risks is part of a larger problem, that of the law's faltering struggle to cope with the onrush of science. It is an old story, but a true one, and becoming more worrisome by the day.") *See also* Michael Livermore, *Catastrophic Risk Review* (LPP Working Paper No. 3-2022, University of Virginia Public Law & Legal Theory Paper Series No. 2022-65 Law & Economics Paper Series No. 2022-21 Sept. 1, 2022) at 16, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4217680 (examining legal, political, and psychological reasons for lawmakers underemphasizing catastrophic risks. "Catastrophic risks are often cross-cutting and are not clearly delegated to specific agencies. … In a constitutional system that is fundamentally grounded in electoral accountability, long-term, global risks will likely be underprioritized. … Human beings have difficulty reasoning about low-probability events and long time horizons; this leads both voters and government officials to neglect catastrophic risks."); SULEYMAN, *supra* note 33, at ch. 13 ("A useful comparison here is climate change. It too deals with risks that are often diffuse, uncertain, temporally distant, happening elsewhere, lacking the salience, adrenaline, and immediacy of an ambush on the savanna—the kind of risk we are well primed to respond to. Psychologically, none of this feels present. Our prehistoric brains are generally hopeless at dealing with amorphous threats like these."). *But see* Global Catastrophic Risk Management Act of 2022, S. 4488, 117th Cong. (2022), https://www.congress.gov/bill/117th-congress/senate-bill/4488/text; *Portman, Peters Introduce Bipartisan Bill to Ensure Federal*

analysis encourages policymakers to dismiss risks that seem improbable or are difficult to quantify.[98] So-called "unknown unknowns" are, by definition, impossible to reliably forecast.[99] Many large-scale societal risks from AI fall into this category.[100]

To be clear, none of this suggests that algorithmic black swans *should* be overlooked. On the contrary, contemporary hazard analysis, which adopts a complex systems perspective, aims to investigate and address difficult-to-anticipate events.[101] It suggests that we can in fact take concrete actions to mitigate unpredictable future risks.[102] Just as experts in the public health, climate science, and financial regulation communities attempt to forecast and address novel tail risks, members of the AI community should prepare to confront the emergence of algorithmic black swans.

II. MARKET FAILURE

A. Steaming Ahead

Current and anticipated risks from AI do not arise in a vacuum. They emerge within an intricate web of research culture, commercial incentives, and regulatory design. The predominant goal of AI research today, which is concentrated in several for-profit industry labs,[103] is to improve capabilities

---

Government is Prepared for Catastrophic Risks to National Security, U.S. SENATE COMMITTEE ON HOMELAND SECURITY & GOVERNMENTAL AFFAIRS (June 24, 2022), https://www.hsgac.senate.gov/media/minority-media/portman-peters-introduce-bipartisan-bill-to-ensure-federal-government-is-prepared-for-catastrophic-risks-to-national-security-. *See also infra* note 340 (discussing proposed revisions to Circular A-4).

[98] *See* Daniel A. Farber, *Uncertainty*, 99 GEO. L.J. 901, 904 (2011) ("Economic modeling and policy analysis are often based on the assumption that extreme harms are highly unlikely … [This] allow[s] extreme risks to be given relatively little weight.") *See also infra* Part V.E (discussing the challenges of quantifying large-scale societal risks).

[99] *See* Donald Rumsfeld, Sec'y, Dep't of Def., DoD News Briefing (Feb. 12, 2002) ("there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones."). For discussion of "unknown unknowns" in AI safety, see Thomas G. Dietterich, *Steps Toward Robust Artificial Intelligence* (Feb. 14, 2016), 30TH AAAI CONF. ON AI, http://videolectures.net/aaai2016_dietterich_artificial_intelligence/.

[100] *See* Kaminski, *supra* note 36, at 153.

[101] *See* LEVESON, *supra* note 93, at 3–6.

[102] *See, e.g.*, Edgar W. Jatho, Logan O. Mailloux, Eugene D. Williams, Patrick McClure & Joshua A. Kroll, *Concrete Safety for ML Problems: System Safety for ML Development and Assessment*, ARXIV (Feb. 6, 2023), https://arxiv.org/abs/2302.02972.

[103] Only exceptionally well-capitalized firms have the resources to carry out frontier AI research. *See* Nathan Benaich & Ian Hogarth, *State of AI Report 2022* at 82 (Oct. 11, 2022),

and performance. Far less attention is given to the technology's potential to cause harm.[104] For example, only around two percent of research papers at the leading AI conference relate to safety.[105] As a result, progress on capabilities continues to outpace progress on safety.[106] According to one researcher at OpenAI: "the capabilities of neural networks are currently advancing much faster than our ability to understand how they work."[107]

What accounts for this imbalance? One explanation is cultural.[108] The prevailing culture in the field of AI, as in the tech sector more broadly, is overwhelmingly technopositive. It advocates unrelenting technical progress, rather than countervailing social or ethical considerations.[109] Other scientific fields, however, are more balanced. In medical research, for instance, analyzing side effects is an accepted, and indeed mandatory, step in developing new treatments. Computer scientists, by contrast, view their work

---

https://www.stateof.ai/2022 ("The compute requirements for large-scale AI experiments has [sic.] increased >300,000x in the last decade. … If the AI community is to continue scaling models, this chasm of "have" and "have nots" creates significant challenges for AI safety"); Nathan Benaich, *State of AI Report 2023* at 75–75, 79 (Oct. 12, 2023), https://www.stateof.ai/ (discussing industrial scale compute clusters and partnerships between AI developers and compute providers and chip manufacturers). *See also* Ganguli et al., *supra* note 24, at 1754–55 (describing the barriers to entry in developing and deploying large models).

[104] *See* Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan & Michelle Bao, *The Values Encoded in Machine Learning Research*, PROC. 2022 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 173 (2022) (empirically studying the values implicit in 100 highly cited machine learning papers). *See also* SULEYMAN, *supra* note 33, at ch. 8.

[105] Dan Hendrycks & Thomas Woodside, *A Bird's Eye View of the ML Field*, ALIGNMENT FORUM (May 8, 2022), https://www.alignmentforum.org/posts/AtfQFj8umeyBBkkxa/a-bird-s-eye-view-of-the-ml-field-pragmatic-ai-safety-2.

[106] *See* Jacob Steinhardt, *AI Forecasting: One Year In*, BOUNDED REGRET (Jul. 3, 2022), https://bounded-regret.ghost.io/ai-forecasting-one-year-in/ ("Overall, progress on machine learning capabilities … was significantly faster than what forecasters expected, while progress on robustness … was somewhat slower than expected.").

[107] *See* Richard Ngo, *The Alignment Problem from a Deep Learning Perspective* at 2 (Aug. 2022) (manuscript on file with author).

[108] *See generally* David Manheim, *Building a Culture of Safety for AI: Perspectives and Challenges* (June 28, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4491421; Hendrycks, Mazeika & Woodside, *supra* note 14, at 25–31. *See also* Sharon Cop & Tal Z. Zarsky, *When Software Meets the Road: Responsibility for Defective Smart Cars in the MVP Era*, 57 GA. L. REV. 1713, 1740–48 (2023) (discussing the safety implications of adopting "agile" and "minimal viable product" software design processes).

[109] *See* Dan Hendrycks & Thomas Woodside, *A Bird's Eye View*, *supra* note 105 ("Researchers are generally quite technopositive … . Much of this tendency is borrowed from the tech industry, which is famously utopian. Likewise, many act as though we must progress towards our predestined technological utopia … . These feelings are amplified in AI because it is perceived to be the next major technological revolution. … Safety and value alignment are generally toxic words …").

through "rose-colored glasses."[110] AI researchers, in particular, often develop and deploy products without considering their societal impact or installing appropriate safeguards.

But these dynamics may be changing. Recent years have seen growing interest and investment in AI safety. For example, leading industry labs employ teams of researchers dedicated to improving the safety of AI systems, as well as teams focused on studying the social and ethical impact of these systems.[111] Meanwhile, the establishment of new academic institutions, independent organizations, and startups has significantly increased the number of people, amount of funding, and volume of research directed toward improving AI safety.[112] These resources, however, still pale in comparison to the resources dedicated to advancing the raw capabilities and performance of automated systems.[113]

The effectiveness of some safety-oriented initiatives is also questionable. For example, since 2020 authors at the most prestigious AI conference, NeurIPS, have been required to submit "broader impact statements" (or similar documents) that address the societal ramifications of their research.[114] While it was hoped that this initiative would surface material safety concerns, this did not pan out. A study interviewing leading AI researchers found that the broader impact statements "have not seriously confronted the issue of the proliferation of dangerous technology."[115] In fact, they may even serve as

---

[110] Brent Hecht et al., *It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process*, ACM FUTURE OF COMPUTING ACADEMY (Mar. 29, 2018), https://perma.cc/K22T-5DFU. *See also* Jeanna Matthews, *Embracing Critical Voices*, 65 COMM. ACM 7 (2022) (critiquing the computer science community's attitude toward the potential negative impacts of technology).

[111] Examples include Google DeepMind's alignment team and OpenAI's safety team.

[112] Academic institutions include UC Berkeley's Center for Human-Compatible AI, and groups at MIT, NYU, Carnegie Mellon University, and the University of Cambridge. Independent organizations include the Alignment Research Center, Redwood Research, and Center for AI Safety. Startups include Anthropic, Conjecture, and Ought.

[113] *See, e.g.*, Benaich & Hogarth, *supra* note 103, at 98 (finding that the number of AI safety researchers "is still orders of magnitude fewer researchers than are working in the broader field, which itself is growing faster than ever" and that "safety funding still trails behind resources for advanced capabilities research").

[114] *See NeurIPS 2020 Call for Papers*, NEURIPS, https://nips.cc/Conferences/2020/CallForPapers; Davide Castelvecchi, *Prestigious AI Meeting Takes Steps to Improve Ethics of Research*, 589 NATURE 12 (2021). The requirement, however, was watered down in 2021, in favor of a checklist document. *See* Carolyn Ashurst et al., *AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements*, PROC. 2022 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 2047, 2047 (2022).

[115] Toby Shevlane, *The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation* at 2 (Apr. 2022) (unpublished D.Phil. dissertation, University of Oxford) (on file with author) [hereinafter Shevlane, *Governing Artefacts*].

window-dressing or ethics-washing,[116] concealing the more consequential harms that could arise from unsafe AI systems. The study concludes that "ethical review is not really a tool for filtering out harmful papers, but rather is a forum for incentivising researchers to change what they *write* in their papers."[117]

Interestingly, some industry-led safety initiatives show greater promise. For instance, several leading large language model developers released a document describing "best practices for deploying language models."[118] The recommendations canvas a wide range of risks and propose concrete mitigation strategies, which some of the organizations proceeded to implement.[119] Clearly, some industry labs are engaging with safety concerns. The problem, however, is that best practices and other deployment-focused frameworks primarily target the immediate risks from current AI systems. Far less attention is directed toward longer-term and larger-scale societal risks, including risks that will arise from new AI technologies and applications. To understand why these risks are overlooked we need to consider broader structural factors.

B. Brinkmanship

Leading AI labs face significant pressure to outpace their competitors. These companies are typically motivated by a combination of economic, scientific, and prestige-related incentives to build AI systems that exhibit state-of-the-art performance.[120] Accelerating development and deployment, and gaining a first-mover advantage, can be particularly valuable. For example, following its release of GPT-3 in 2020 and, two years later,

---

[116] *See, e.g.*, Elettra Bietti, *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*, PROC. 2020 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 210 (2020).

[117] Shevlane, *Governing Artefacts*, *supra* note 115, at 8 (emphasis added).

[118] *See* OpenAI, *Best Practices for Deploying Language Models*, OPENAI (June 2, 2022), https://openai.com/blog/best-practices-for-deploying-language-models/ (describing a joint governance initiative of Cohere, OpenAI, and AI21 Labs).

[119] *See* Miles Brundage, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike & Pamela Mishkin, *Lessons Learned on Language Model Safety and Misuse*, OPENAI (Mar. 3, 2022), https://openai.com/blog/language-model-safety-and-misuse/.

[120] *See* Ganguli et al., *supra* note 24, at 1754. *See also* Jack Clark & Gillian K. Hadfield, *Regulatory Markets for AI Safety*, ARXIV at 2 (Dec. 11, 2019), https://arxiv.org/abs/2001.00078 ("Companies competing in markets have an incentive to build AI faster than their competitors, and … assuring the safety of large-scale machine learning-driven systems appears to be both costly and difficult; slowing that process down while encouraging an environment for investment to ensure safe development is a collective action problem that regulation is needed to address.").

ChatGPT, OpenAI became the clear frontrunner in language model technology. Google, its AI research subsidiary, DeepMind, and Meta scrambled to catch up.[121]

These competitive market dynamics can certainly produce prosocial outcomes. The faster the pace of AI progress, the faster users can deploy the technology in prosocial applications. For instance, shortly after its release, GPT-3 was used to assist consumers in understanding the terms of standard form contracts.[122] The problem, however, is that the market dynamics of the AI ecosystem incentivize speed, rather than safety.[123]Allan Dafoe, a research scientist at Google DeepMind, describes the danger as follows: "race settings—where actors perceive large gains from relative advantage—can induce actors to cut corners, exposing the world to risks that they would otherwise prudently avoid."[124]

This brinkmanship affects organizational decisions at every stage in the AI value chain. Refraining from building or deploying systems that exhibit the most impressive performance can damage a company's reputation and bottom line.[125] Allocating resources toward improving safety may come at the expense of other more lucrative investments. And, even if an organization successfully develops effective safety mechanisms, it may be reluctant to implement these if they hamstring a system's performance or profitability.[126]

---

[121] *See* Ganguli et al., *supra* note 24, at 1755; Kevin Roose, *How ChatGPT Kicked Off an A.I. Arms Race*, N.Y. TIMES (Feb. 3, 2023), https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html; Madhumita Murgia, *Google's DeepMind-Brain Merger: Tech Giant Regroups for AI Battle*, FIN. TIMES (Apr. 28, 2023), https://www.ft.com/content/f4f73815-6fc2-4016-bd97-4bace459e95e; Will Knight, *Google DeepMind's CEO Says Its Next Algorithm Will Eclipse ChatGPT*, WIRED (June. 26, 2023), https://www.wired.com/story/google-deepmind-demis-hassabis-chatgpt/.

[122] *See* Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. 83 (2022); Noam Kolt, *Predicting Consumer Contracts*, 37 BERKELEY TECH. L.J. 71 (2022).

[123] *See* Hendrycks, Mazeika & Woodside, *supra* note 14, at 17–18; Amanda Askell, Miles Brundage & Gillian Hadfield, *The Role of Cooperation in Responsible AI Development*, ARXIV at 8–10 (Jul. 10, 2019), https://arxiv.org/abs/ 1907.045342. *See also* Nitasha Tiku, Gerrit De Vynck & Will Oremus, *Big Tech Was Moving Cautiously on AI. Then Came ChatGPT*, WASH. POST (Jan. 27, 2023), https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/.

[124] Allan Dafoe, *AI Governance: Overview and Theoretical Lenses*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (Justin B. Bullock et al. eds., forthcoming). *See also* Stephen Cave & Seán S ÓhÉigeartaigh, *An AI Race for Strategic Advantage: Rhetoric and Risks*, PROC. 2018 AAAI /ACM CONF. AI, ETHICS & SOC'Y 36 (2018).

[125] *See* Askell, Brundage & Hadfield, *supra* note 123; Clark & Hadfield, *supra* note 120.

[126] This is sometimes described as an "alignment tax." *See*, *e.g.*, Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment*, ARXIV at 11–14 (Dec. 9, 2021), https://arxiv.org/abs/2112.00861. For a concrete illustration of this tradeoff, see Alexander Pan et al., *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark*, 202 PROC. MACH. LEARNING.

Faced with these incentives, AI developers will have to choose between increasing their risk tolerance and decreasing their investment in research. In a dangerous "race to the bottom," safety-conscious firms could fall behind as cutting-edge AI development becomes dominated by companies with the greatest risk appetite and least concern for the societal impact of the technologies they create.[127]

C. Externalities

Fierce competition between commercial AI labs is not the only explanation for the under-investment in safety. The market failure can also be explained by *who* bears the costs of unsafe automated systems. To make this concrete, consider the impact of AI-powered code generation tools (such as GitHub Copilot), which assist human programmers in writing software. As discussed above, these systems can produce code that contains bugs and security vulnerabilities.[128]

Who suffers the resulting harm? To begin with, software engineers who use defective code generation tools suffer harm because these tools decrease the quality of the software produced. However, it is the end-users of the resulting software who, despite never using code generation tools themselves, suffer most. They stand to bear the cost of dangerous bugs and vulnerabilities, yet have little recourse against the company who built the defective code generation tool in the first place. This indirect and diffuse harm is, like carbon emissions, a negative externality.[129] Producers of the externality, namely, developers of code generation tools, have limited incentive to mitigate the harm or redress the losses incurred.

This type of market failure, common to many AI products and services, is aggravated by several factors. First, because the harm is caused indirectly, end-users who suffer losses may find it difficult to attribute liability to the developer of the code generation tool.[130] After all, software engineers who fail to adequately vet the security of the code may themselves be liable. Second, because the harm is diffuse, there arises a collective action problem: end-users who suffer comparatively small losses individually but large losses in aggregate may struggle to coordinate in taking action against the developer

---

RES. 26837 (2023).

[127] *See* Stuart Armstrong et al., *Racing to the Precipice: A Model of Artificial Intelligence Development*, 31 AI & SOC. 201 (2016).

[128] *See supra* Part I.B.

[129] *See* Askell, Brundage & Hadfield, *supra* note 123, at 6.

[130] *See* Kaminski, *supra note* 36, at 119; Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. L. & TECH. 353, 356 (2016); Anat Lior, *The AI Accident Network: Artificial Intelligence Liability Meets Network Theory*, 95 TUL. L. REV. 1103, 1117–18 (2021).

of the code generation tool.[131] Third, in some cases it may in fact be impossible for a single company (or insurer) to compensate for the harm caused by a code generation tool.[132] For example, the economic damage caused by critical vulnerabilities in widely used software may be incalculable.[133]

The upshot of this analysis is that organizations building AI systems bear only a fraction of the costs of harms they may cause, especially in the case of black swans. Without intervention, these organizations are unlikely to sufficiently increase their investment in safety measures and risk mitigation.

* * * * *

For completeness, two additional factors contribute to the pervasive under-investment in AI safety. The first relates to the technical abilities of the users of AI systems. Continuing with the case of AI-powered code generation tools, software engineers may be unable to evaluate the security of the code these tools produce, such that they will not adjust their willingness to pay for them.[134] The second factor relates to the technical abilities of AI developers. Even if software engineers demand safer code generation tools, AI developers may be unable to meet this demand.[135] Software engineers, aware of this limitation, may refrain from demanding safety guarantees in the first place—further eroding the motivation of AI developers to increase their investment in safety.

---

[131] Data-driven technologies frequently give rise to such collective action problems. *See, e.g.*, Jef Ausloos, Jill Toh & Alexandra Giannopoulou, *The Case for Collective Action Against the Harms of Data-driven Technologies*, ADA LOVELACE INST. (Nov. 23, 2022), https://www.adalovelaceinstitute.org/blog/collective-action-harms/. *See also* Rebecca Crootof & BJ Ard, *Structuring Techlaw*, 34 HARV. J.L. & TECH. 348, 382 (2021).

[132] *See* Askell, Brundage & Hadfield, *supra* note 123, at 6. *See also* Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715, 1738 (2015) ("Externality producers may also fail to take full account of future liabilities if they expect to be judgment-proof by the time enforcement occurs. Prior commentators describe this as a problem of liquidity, that is, the producer lacks the cash to cover its penalty, and cannot borrow enough money to pay.").

[133] *See, e.g.*, Gregory S. Gaglione, Jr., Comment, *The Equifax Data Breach: An Opportunity to Improve Consumer Protection and Cybersecurity Efforts in America*, 67 BUFF. L. REV. 1133, 1154–1166 (2019) (discussing the ramifications of a security vulnerability in a widely used software framework that led to a breach in which the financial records of over 150 million consumers were compromised).

[134] *See* Askell, Brundage & Hadfield, *supra* note 123, at 4.

[135] *Id.*

III. THE EVOLVING LEGAL LANDSCAPE

As we can see, market forces cannot address the full range of societal challenges presented by AI. The case for robust policy intervention is strong, as reflected in the proliferation of proposals for regulating AI. In the United States, the number of AI-related state bills increased from five in 2015 to sixty in 2022.[136] These proposals add to, and sometimes modify, existing rules in tort law, consumer law, administrative law, and a variety of sector-specific regulations.[137]

While governments in many countries, including China,[138] Canada,[139] and the United Kingdom,[140] have proposed national plans for regulating AI, this Article focuses on arguably the two most important jurisdictions: the United States and the European Union.[141] While the United States is home to the world's largest AI market and research ecosystem,[142] the European Union is a regulatory "superpower" that exercises outsized influence on the rules and standards in global markets.[143] Given the EU's track record in passing field-defining laws for emerging technologies, such as the General Data

---

[136] *See* Maslej et al., *supra note* 54, at 274. Notably, a growing fraction of these bill has been passed into law. For summary of recent AI-related state bills, see *Artificial Intelligence 2023 Legislation*, NATIONAL CONFERENCE OF STATE LEGISLATURES (Sept. 27, 2023), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation.

[137] *See*, *e.g.*, Calo, *Primer*, *supra* note 5, at 427–31.

[138] *See*, *e.g.*, Matt Sheehan, *China's AI Regulations and How They Get Made*, CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE (July 2023), https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

[139] *See infra* note 278–79 (discussing Canada's proposal for an AI and Data Act, which is loosely modeled on the EU AI Act).

[140] *See Establishing a Pro-Innovation Approach to Regulating AI*, UK GOVERNMENT (Jul. 20, 2022), https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement;

[141] The UK, however, appears to be an increasingly important jurisdiction in terms of regulation that targets large-scale risks from AI. *See Frontier AI Taskforce: First Progress Report*, UK DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY (Sept. 7, 2023), https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report; *AI Safety Summit: Introduction*, UK DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY (Sept. 25, 2023), https://www.gov.uk/government/publications/ai-safety-summit-introduction; *National AI Strategy* (Dec. 18, 2022), https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version. *But see supra* note 140.

[142] *See* Maslej et al., *supra* note 54, at 189 (showing that the United States continues to lead the world in AI-related private investment).

[143] *See* ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD xiii (2020); Anu Bradford, *The Brussels Effect*, 107 NW. U. L. REV. 1, 66–67 (2012). *See also infra* Part III.A.3 (discussing the Brussels effect of EU AI regulations).

Protection Regulation (GDPR),[144] and its early efforts to enact comprehensive legislation for automated systems, this survey begins with the major EU proposals for AI regulation.

A. European Union

1. EU AI Act

The European Commission's proposal for an Artificial Intelligence Act (the EU AI Act),[145] the first version of which was published in April 2021, is the first attempt to comprehensively regulate AI systems in a major jurisdiction.[146] The proposal, which is expected to be passed into law,[147] covers a wide range of AI technologies and applications.[148] Once in effect, the Act will apply to all EU Member States, barring them from passing domestic laws that conflict with provisions of the Act.[149] The Act is anticipated to transform AI regulation in much the same way as the GDPR

[144] Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[145] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM (2021) 206 final (Apr. 21, 2021) [hereinafter EU AI Act]. Unless otherwise indicated, all references to the EU AI Act are to the General Approach of the European Council (Nov. 25, 2022) adopted on December 6, 2022, available at https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf. Negotiations to finalize the text of the Act are currently underway between the European Parliament, the European Council, and the European Commission. *See Legislative Train Schedule*, EUROPEAN PARLIAMENT (Sept. 20, 2023), https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence. For a comparison of the texts proposed by the respective institutions, see *Trilogue Mandates* (June 20, 2023) https://artificialintelligenceact.eu/wp-content/uploads/2023/08/AI-Mandates-20-June-2023.pdf. For a summary of the European Parliament's position, which was adopted on June 14, 2023, see *Parliament's Negotiating Position on the Artificial Intelligence Act*, EUROPEAN PARLIAMENT (June 2023) https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/747926/EPRS_ATA(2023)747926_EN.pdf.

[146] *See* Michael Veale & Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, 22 COMPUT. L. REV. INT'L 97, 112 (2021).

[147] *See* European Parliament, *EU AI Act: First Regulation on Artificial Intelligence*, EUROPEAN PARLIAMENT NEWS (June. 8, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence; *Legislative Train Schedule*, *supra* note 145.

[148] *See* EU AI Act art. 3(1) (proposing a broad definition of AI systems).

[149] *See* Kaminski, *supra* note 36, at 128.

transformed data privacy law.[150]

The AI Act adopts a risk-based approach, classifying the uses of AI according to the potential harm they may cause. Certain uses of AI, such as social scoring and real-time remote biometric surveillance, are prohibited.[151] Other uses, such as AI systems deployed in hiring contexts or integrated into medical devices, are deemed "high-risk" and subject to conformity assessments that include, among other things, transparency requirements, human oversight, and detailed record-keeping.[152] Meanwhile, uses that are considered to pose more limited risks are subject to less onerous transparency requirements.[153]

Notably, the Act establishes significant penalties for infringement: fines of up to 6% of a firm's global revenue or €30 million (whichever is higher) in respect of prohibited uses; and fines of up to 4% of a firm's global revenue or €20 million (whichever is higher) in respect of a failure to comply with certain requirements applicable to high-risk uses and limited risk uses.[154]

Finally, despite the apparent comprehensiveness of the Act—which contains eighty-five articles and runs over 200 pages—it leaves open important issues. Details of certain requirements in the Act will only be specified in "implementing acts" at some point in the future.[155] Meanwhile, technical standards, compliance with which gives rise to a presumption of conformity with the Act,[156] have not yet been written. In fact, this task has been outsourced to private standard-setting organizations.[157] Despite not

---

[150] *See* Lilian Edwards, *Regulating AI in Europe: Four Problems and Four Solutions*, ADA LOVELACE INST. at 2 (Mar. 2022), https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/. The GDPR itself also contains several provisions that govern the use of AI. *See* GDPR art. 22 (automated individual decision-making), arts.13–15 (notification and access rights). On the relationship between the GDPR and the EU AI Act, see Sebastião Barros Vale, *GDPR and the AI Act Interplay*, FUTURE OF PRIVACY FORUM (Nov. 3, 2022), https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report/. On the relationship between the EU AI Act and other EU laws, see Artur Bogucki, Alex Engler, Clément Perarnaud & Andrea Renda, *The AI Act and Emerging EU Digital* Acquis, CENTRE FOR EUR. POL'Y STUD. (Sept. 2, 2022), https://www.ceps.eu/ceps-publications/the-ai-act-and-emerging-eu-digital-acquis/.

[151] EU AI Act tit. II.

[152] *Id.* tit. III.

[153] *Id.* tit. IV. *See also* EU AI Act tit. IX (establishing voluntary codes of conduct for minimal risk uses).

[154] EU AI Act art. 71(3)–(4). *Compare* GDPR art. 83(4)–(5) (establishing fines of up to 2% or 4% of global revenue, or €10 million or €20 million, depending on the infringement).

[155] This, however, is not uncommon for EU legislation. *See Implementing and Delegated Acts*, EURO. COMMISSION, https://commission.europa.eu/law/law-making-process/adopting-eu-law/implementing-and-delegated-acts_en.

[156] *Id.* art. 40.

[157] *See* European Commission, A Notification under Article 12 of Regulation (EU) No 1025/20121 (Dec. 2, 2022) (proposing a standardization request to the European Committee

appearing in the legislative text, these organizations could ultimately determine the real-world impact of the Act.[158]

2. EU AI Liability Directive

In September 2022, the European Commission proposed a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (the EU AI Liability Directive).[159] The Directive aims to complement the EU AI Act by facilitating civil claims in respect of harms caused by AI systems.[160] The Act and the Directive are two sides of the same coin: the former is intended to prevent harm from occurring; the latter is intended to provide compensation if harm occurs.[161]

---

for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC), two private standard-setting organizations).

[158] *See* Veale & Zuiderveen Borgesius, *supra* note 146, at 105 ("standardisation is arguably where the real rule-making in the … Act will occur"). *See id.* (contextualizing this practice within the EU's New Legislative Framework and suggesting that "The practice of delegating rule-making to bodies governed by private law such as CEN/CENELEC is controversial and sits on increasingly shaky legal ground."). *See also* Hadrien Pouget, *The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist*, LAWFARE (Jan. 12, 2023), https://www.lawfareblog.com/eus-ai-act-barreling-toward-ai-standards-do-not-exist (suggesting that compliance with EU AI Act standards may be technically unfeasible).

[159] European Commission, Proposal for a Directive of the European Parliament and of the Council on Adapting Noncontractual Civil Liability Rules to Artificial Intelligence COM (2022) 496 final (Sept. 28, 2022) [hereinafter EU AI Liability Directive].

[160] Notably, upon the publication of the EU AI Act in 2021, many commentators were surprised, and disappointed, that (unlike the GDPR) the AI Act did not contain a mechanism for private enforcement. *See, e.g.*, Veale & Zuiderveen Borgesius, *supra* note 146, at 111.

[161] For further discussion (and criticism) of the EU AI Liability Directive, see Philipp Hacker, *The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future*, 51 COMPUT. L. & SEC. REV. (2023); Marta Ziosi, Jakob Mökander, Claudio Novelli, Federico Casolari, Mariarosaria Taddeo & Luciano Floridi, *The EU AI Liability Directive: Shifting the Burden From Proof to Evidence* (Working Paper, June 21, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4470725; Claudia Prettner, *FLI Position Paper on AI Liability*, FUTURE OF LIFE INST. (Nov. 2022), https://futureoflife.org/wp-content/uploads/2022/11/FLI_AI_Liability_Position_Paper.pdf; Orian Dheu, Jan De Bruyne & Charlotte Ducuing, *The European Commission's Approach to Extra-Contractual Liability and AI: An Evaluation of the AI Liability Directive and the Revised Product Liability Directive* (Working Paper, Dec. 7, 2022), https://papers.ssrn.com/sol3/ papers.cfm?abstract_id=4295676. For discussion of liability under other EU regulations, see Philipp Hacker, Andreas Engel & Marco Mauer, *Regulating ChatGPT and other Large Generative AI Models*, PROC. 2023 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1112 (2023); Roee Sarel, *Restraining ChatGPT*, HASTINGS L.J. (forthcoming); Philipp Hacker, *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, *in* OXFORD HANDBOOK OF ALGORITHMIC GOVERNANCE AND THE LAW (Ifeoma Ajunwa & Jeremias Adams-Prassl eds., forthcoming).

The Directive, which EU Member states must implement domestically, proposes two mechanisms to assist victims seeking redress in respect of harm caused by AI. First, the Directive establishes a "presumption of causality": if a claimant can demonstrate that the defendant failed to comply with certain requirements under the AI Act, a court will presume that such non-compliance *caused* the relevant harm.[162] Second, the Directive empowers courts to order the disclosure of evidence related to certain civil claims arising from the operation of high-risk AI systems.[163]

3. Brussels Effect

Although the EU AI Act and EU AI Liability Directive apply solely to AI systems and users in the European Union,[164] these regulations are likely to have a global impact. The outsized influence of EU regulation on the rules and standards in global markets—known as the "Brussels Effect"—is a well-document phenomenon.[165] It takes two forms. The *de facto* Brussels Effect involves multinational firms standardizing their production globally in order to comply with EU regulations, such that products manufactured outside of the European Union for non-EU customers will in practice comply with EU regulations.[166] The *de jure* Brussels Effect involves countries outside of the European Union adopting regulations and standards similar to those established inside the European Union.[167]

Observers expect the EU AI Act to exhibit both a *de facto* and *de jure* Brussels effect. The Act, they suggest, will incentivize AI developers outside of the European Union to build systems that comply with EU regulations and prompt countries outside of the European Union to pass regulations that align with the requirements established in the EU AI Act.[168]

---

[162] EU AI Liability Directive art. 4.

[163] *Id.* art. 3.

[164] *See* EU AI Act art. 2 (defining the scope of the Act). EU directives, including the EU AI Liability Directive, apply solely to EU Member States.

[165] *See generally* BRADFORD, *supra* note 143; Bradford, *supra* note 143. A prominent recent example of the Brussels Effect is the impact of the GDPR on data privacy laws outside the European Union. *See* Cedric Ryngaert & Mistale Taylor, *The GDPR as Global Data Protection Regulation?*, 114 AM. J. INT'L L. UNBOUND 5 (2020); Paul M. Schwartz, *Global Data Privacy: The EU Way*, 94 N.Y.U. L. REV. 771 (2019).

[166] *See* Bradford, *supra* note 143, at 6 ("While the EU regulates only its internal market, multinational corporations often have an incentive to standardize their production globally and adhere to a single rule.").

[167] *Id.* ("after these export-oriented firms have adjusted their business practices to meet the EU's strict standards, they often have the incentive to lobby their domestic governments to adopt these same standards in an effort to level the playing field against their domestic, non-export-oriented competitors").

[168] *See* Charlotte Siegmann & Markus Anderljung, *The Brussels Effect and Artificial*

This Brussels Effect may indeed already be underway. Fourteen months after the European Commission published the first draft of the EU AI Act, Canada's parliament introduced the Artificial Intelligence and Data Act,[169] which is loosely modeled on the EU AI Act. For example, the Canadian Artificial Intelligence and Data Act proposes a relatively comprehensive regulatory regime, adopts a risk-based approach, and establishes penalties similar to those set out in the EU AI Act.[170]

A *de jure* Brussels Effect could also occur in the United States. Just as the European Union's GDPR dramatically shaped the development of data privacy law in California,[171] the EU AI Act could inspire new approaches to regulating AI in the United States at both the state and federal level.[172]

B. United States

1. NIST AI Risk Management Framework

In March 2022, the National Institute of Standards and Technology (NIST), an agency of the U.S. Department of Commerce, released the initial draft of its AI Risk Management Framework (RMF).[173] Finalized in January

Intelligence: How EU Regulation Will Impact the Global AI Market, CENTRE FOR THE GOVERNANCE OF AI at 3–5 (Aug. 2022), https://www.governance.ai/research-paper/ brussels-effect-ai; Edwards, *supra* note 150, at 2. *Compare* Alex Engler, *The EU AI Act Will Have Global Impact, but a Limited Brussels Effect*, BROOKINGS INST. (June 16, 2022), https://www.brookings.edu/blog/ techtank/2022/06/14/the-limited-global-impact-of-the-eu-ai-act/; Marco Almada & Anca Radu, *The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy*, GERMAN L.J. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4592006.

[169] House of Commons of Canada, Bill C-27, An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts (or Digital Charter Implementation Act, 2022) (June 16, 2022) (Can.).

[170] There are, however, notable differences between the two proposals. For example, while the EU AI Act applies to both government and private actors, the Canadian proposal largely excludes governmental entities; and, unlike the EU AI Act, the Canadian proposal does not impose any blanket prohibits on certain uses of AI. In addition, the Canadian proposal is considerably less detailed than the EU AI Act. For instance, the definition of a "high-impact system" (§ 5(1)) and associated compliance requirements (§ 8) will only be established in future regulations.

[171] Schwartz, *supra* note 165, at 817 ("The EU had not set up a policy shop in Sacramento, California. It had not lobbied the state legislature or Governor ... Yet, somehow, the ideas of EU data protection made their way to the Golden State.").

[172] *See*, *e.g.*, National AI Commission Act, H.R. 4223, 118th Cong. (2023) (offering support for a risk-based approach, somewhat comparable to that of the EU AI Act).

[173] For information regarding the development of the framework, including public involvement, see *AI Risk Management Framework*, NAT'L INST. STANDARDS & TECH., https://www.nist.gov/itl/ai-risk-management-framework.

2023, the RMF is a voluntary framework that aims to assist organizations in anticipating and addressing risks from AI.[174] Concretely, the RMF resembles an enterprise risk management framework.[175] It aims to cultivate an organizational safety culture in which stakeholders "map" the AI risk landscape and develop methods to appropriately "measure" and "manage" these risks.[176]

To be clear, NIST's framework does not impose any legal obligations.[177] It is a light-touch, "quintessentially American" soft law regulatory tool.[178] That being said, the framework could be highly influential if government or corporate procurement contracts for AI systems were to mandate that vendors comply with the principles set out in the RMF.[179] Depending on the nature and size of these contracts, AI developers may be incentivized to make substantial changes to the products and services they offer. In addition, insurers and courts could expect AI developers to demonstrate compliance with the RMF, much like they expect companies to operate in accordance with NIST cybersecurity frameworks.[180] Finally, elements of the RMF could eventually be integrated into binding sector-specific regulations.[181]

---

[174] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NAT'L INST. STANDARDS & TECH. at 2 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [hereinafter NIST AI Risk Management Framework or NIST AI RMF].

[175] Note, however, that parts of the EU AI Act also incorporate elements of enterprise risk management. *See* Jonas Schuett, *Risk Management in the Artificial Intelligence Act*, ARXIV (Dec. 3, 2022), https://arxiv.org/abs/2212.03109 (discussing Article 9 of the EU Act).

[176] *See* NIST AI RMF, *supra* note 174, at 20.

[177] *See id.* at 2, 7. NIST, it should be clarified, is a non-regulatory agency.

[178] *See* Kaminski, *supra* note 36, at 55 (in earlier SSRN manuscript, dated Oct. 27, 2022) (on file with author).

[179] *See* Louis Au Yeung, *Guidance for the Development of AI Risk and Impact Assessments*, UC BERKELEY CENTER FOR LONG-TERM CYBERSECURITY at 16, 27 (Jul. 2021), https://cltc.berkeley.edu/publication/guidance-for-the-development-of-ai-risk-and-impact-assessments/; Fei-Fei Li, *Governing AI Through Acquisition and Procurement*, STANFORD UNIVERSITY INSTITUTE FOR HUMAN-CENTERED AI (Sept. 14, 2023), https://hai.stanford.edu/sites/default/files/2023-09/Fei-Fei-Li-Senate-Testimony.pdf. *See also* Darrell M. West, *California Charts the Future of AI*, BROOKINGS (Sept. 12, 2023), https://www.brookings.edu/articles/california-charts-the-future-of-ai/ (explaining that a recent California executive order "leverage[s] the state's procurement power to promote trustworthy AI principles.").

[180] *See* Anthony M. Barrett, Dan Hendrycks, Jessica Newman & Brandie Nonnecke, *Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks*, ARXIV at 5 (Sept. 7, 2022), https://arxiv.org/abs/2206.08966. *See also* Kaminski, *supra* note 36, at 162 (explaining that the NIST AI RMF is modeled on the NIST Cybersecurity Framework that was introduced in 2014).

[181] *See* Barrett, Hendrycks, Newman & Nonnecke, *supra* note 180.

2. White House AI Bill of Rights

Another important federal initiative for regulating AI is the Blueprint for an AI Bill of Rights, released by the White House Office of Science and Technology Policy in October 2022.[182] Despite its constitution-evoking title, the document is best described as aspirational. The blueprint—which states that it is "*non-binding and does not constitute U.S. government policy*"[183]—offers a broad contextualization of the risks from AI, and outlines five principles for guiding the design, use, and regulation of AI systems. These include: (i) protection from unsafe and ineffective systems; (ii) prevention of algorithmic discrimination; (iii) protection of data privacy; (iv) disclosure and explanation of the use of AI systems; and (v) access to human alternatives in place of AI systems.[184]

The central problem with the blueprint is not that it lacks "teeth", but that it does not meaningfully explore how the five principles it enshrines will be implemented in practice.[185] Statements like "some of the additional protections proposed in this framework would require new laws to be enacted or new policies and practices to be adopted" offer little clarity.[186] The pathway forward, on a federal level, is likely to involve a combination of executive orders, sector-specific regulations, and other federal agency actions. For example, the Federal Trade Commission, Equal Employment Opportunity Commission, and Consumer Financial Protection Bureau are expected to develop regulations and practices in their respective domains.[187]

---

[182] White House Office of Sci. & Tech. Policy, *Blueprint for an AI Bill of Rights: Making Automated Systems Work* (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [hereinafter White House AI Bill of Rights].

[183] *Id.* at 2 (the legal disclaimer further clarifies that the blueprint "does not supersede, modify, or direct an interpretation of any existing statute, regulation, policy, or international instrument. It does not constitute binding guidance for the public or Federal agencies and therefore does not require compliance with the principles described herein.").

[184] *Id.* at 5–7. The principles are elaborated upon in a 62-page "technical companion."

[185] *See* Khari Johnson, *Biden's AI Bill of Rights Is Toothless Against Big Tech*, WIRED (Oct. 4, 2022), https://www.wired.com/story/bidens-ai-bill-of-rights-is-toothless-against-big-tech/. *Compare* Alex Engler, *The AI Bill of Rights Makes Uneven Progress on Algorithmic Protections*, LAWFARE (Oct. 7, 2022), https://www.lawfareblog.com/ai-bill-rights-makes-uneven-progress-algorithmic-protections; Emmie Hine & Luciano Floridi, *The Blueprint for an AI Bill of Rights: In Search of Enaction, at Risk of Inaction*, 33 MINDS & MACH. 285, 286 (2023).

[186] White House AI Bill of Rights, *supra* note 182, at 8.

[187] *See* White House Office of Sci. & Tech. Policy, *Biden-Harris Administration Announces Key Actions to Advance Tech Accountability and Protect the Rights of the American Public* (Oct. 2, 2022), https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/. The White House also secured voluntary commitments from several leading AI companies. *See* White House Office of Sci.

Of course, the content and timing of these measures remain to be seen.[188]

3. Legislative Proposals

Alongside various non-binding instruments for regulating AI, several bills and legislative frameworks have been proposed at the federal level. These include the Algorithmic Accountability Act, first introduced in 2019, which seeks to impose mandatory obligations on certain uses of automated systems.[189] The bill, revised versions of which were introduced in February 2022 and September 2023, would require the Federal Trade Commission (FTC) to promulgate regulations requiring companies to conduct impact assessments of automated decision processes and publish annual reports based on these impact assessments,[190] The scope of the Algorithmic Accountability Act, however, is far more limited than that of the EU AI Act. It applies only to companies over which the FTC has jurisdiction, which excludes public agencies, banks, air carriers, and other high-impact users of automated systems.[191] In addition, unlike the EU AI Act, the Algorithmic Accountability Act does not impose stringent conditions on high-risk uses of AI, much less prohibit certain uses outright.

Another notable proposal is the National AI Commission Act,[192] introduced with bipartisan support in June 2023.[193] The bill aims to establish

---

& Tech. Policy, *Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (Jul. 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; White House Office of Sci. & Tech. Policy, *Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI* (Sept. 12, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[188] Enforcement is also a key challenge. *See, e.g.*, Christie Lawrence, Isaac Cui & Daniel Ho, *The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies*, PROC. 2023 AAAI /ACM CONF. AI, ETHICS & SOC'Y 606 (2023) (finding that federal agencies implemented only a fraction of existing AI governance requirements).

[189] Algorithmic Accountability Act of 2023, S. 2893, 118th Cong. (2023); Algorithmic Accountability Act of 2023, H.R. 5628, 118th Cong. (2023) [hereinafter Algorithmic Accountability Act]. For further analysis of the bill, see Kaminski, *supra* note 36, at 134–36; Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J. L. & TECH. 117, 146–52 (2021) (discussing the 2019 version of the bill).

[190] *See* Algorithmic Accountability Act §§ 3–6.

[191] *Id*. at §§ 2 (defining "covered entity").

[192] National AI Commission Act, H.R. 4223, 118th Cong. (2023).

[193] *See H.R. 4223: National AI Commission Act*, GOVTRACK (accessed Oct. 3, 2023);

a National AI Commission that would "work to ensure … that through regulation the United States is mitigating the risks and possible harms of artificial intelligence" and "takes a leading role in establishing necessary, long-term guardrails to ensure that artificial intelligence is aligned with values shared by all Americans".[194] The Commission would also "build upon previous Federal efforts and international best practices and efforts to develop a *binding risk-based approach* to regulate and oversee artificial intelligence applications through identifying applications with unacceptable risks, high or limited risks, and minimal risks."[195] While the bill's approach might appear to align with the EU AI Act, it is worth recalling that the bill only proposes the establishment of a Commission, but not concrete or binding EU-style regulatory and enforcement mechanisms.

Finally, several prominent lawmakers have proposed legislative frameworks for regulating AI. These include the SAFE Innovation Framework, proposed by Senate Majority Leader Chuck Schumer in June 2023.[196] The framework describes five main policy objectives for governing AI but offers little practical guidance on how these would be accomplished.[197] In September 2023, Senators Hawley (R-MO) and Blumenthal (D-CT) published a more detailed bipartisan framework for regulating AI.[198] The framework proposes establishing a licensing regime for general purpose AI systems, to be administered by an independent oversight body that would have the authority to conduct audits of AI developers.[199] In addition, the framework proposes holding AI developers liable for certain harms caused by AI, including by waiving immunity under Section 230 of the

---

https://www.govtrack.us/congress/bills/118/hr4223; *Reps Lieu, Buck, Eshoo and Sen Schatz Introduce Bipartisan, Bicameral Bill to Create a National Commission on Artificial Intelligence*, TED LIEU (June 20, 2023), https://lieu.house.gov/media-center/press-releases/reps-lieu-buck-eshoo-and-sen-schatz-introduce-bipartisan-bicameral-bill.

[194] National AI Commission Act § 3(g)(1).

[195] *Id.* at § 3(g)(4) (emphasis added).

[196] *SAFE Innovation Framework*, SENATE MAJORITY LEADER CHUCK SCHUMER (June 21, 2023) https://www.democrats.senate.gov/imo/media/doc/schumer_ai_framework.pdf

[197] Some additional details are available in Schumer's remarks. *See Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS, Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS*, CENTER FOR STRATEGIC & INT'L STUDIES (June 21, 2023), https://www.csis.org/analysis/sen-chuck-schumer-launches-safe-innovation-ai-age-csis.

[198] Senators Richard Blumenthal & Josh Hawley, *Bipartisan Framework for U.S. AI Act*, https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf [hereinafter Blumenthal-Hawley Framework]. For analysis of the proposal, see Tessa Baker, *Blumenthal and Hawley's U.S. AI Act Framework: CSET's Perspective and Contributions* (Sept. 19, 2023), GEORGETOWN CENTER FOR SECURITY & EMERGING TECHNOLOGY https://cset.georgetown.edu/article/blumenthal-and-hawleys-u-s-ai-act-framework-csets-perspective-and-contributions/.

[199] *Id.*

Communications Act of 1934 for certain harms relating to AI.[200] The framework also proposes that the outputs of AI systems be "watermarked" so as to enable users to identify that they are AI-generated.[201]

IV. GOVERNANCE GAPS

Focusing on laws and policy instruments that have already been enacted (or are likely to be enacted),[202] the European Union and the United States have taken very different approaches to regulating AI. While the EU has opted for binding horizontal regulation, the U.S. is largely moving toward voluntary sector-specific governance.[203] However, most of the EU and U.S. proposals share one thing in common: they primarily target the immediate risks from AI, rather than broader, longer-term risks.[204] Regulatory efforts on both sides of the Atlantic overlook the risk of algorithmic black swans, either neglecting catastrophic tail risks altogether or adopting governance mechanisms with problematic gaps. The following section examines three of the most salient black swan risks that EU and U.S. regulatory proposals fail to address: high-impact accident risks from general purpose AI systems; the uncontrolled proliferation and malicious use of AI systems; and applications of AI that could cause long-term systemic harm to social and political institutions.

---

[200] *Id. See also* A Bill to Waive Immunity Under Section 230 of the Communications Act of 1934 for Claims and Charges Related to Generative Artificial Intelligence, S. 1993, 118th Cong. (2023) (sponsored by Senators Hawley and Blumenthal).

[201] *Bipartisan Framework for U.S. AI Act*, *supra* note 192. Corresponding bills with bipartisan support have been proposed. *See* AI Labeling Act of 2023, S. 2691, 118th Cong. (2023); AI Disclosure Act of 2023, H.R. 3831, 118th Cong. (2023). For discussion of technical methods and limitations of watermarks, see John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers & Tom Goldstein, *A Watermark for Large Language Models*, 202 PROC. MACH. LEARN. RES. 17061 (2023); John Kirchenbauer et al., *On the Reliability of Watermarks for Large Language Models*, ARXIV (June 7, 2023), https://arxiv.org/abs/2306.04634.

[202] The EU AI Act is likely to become law. *See supra* note 145.

[203] One notable exception is the Blumenthal-Hawley Framework. However, no bills have (at the time of writing) been introduced to implement the framework.

[204] A possible exception in the U.S. is the National AI Commission Act § 3(g)(1) (explicitly referring to the need for "long-term guardrails"). *See also National Artificial Intelligence Research and Development Strategic Plan 2023 Update*, NAT'L SCI. & TECH. COUNCIL SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE at 16 (May 2023), https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.

A. General Purpose Systems

Recent progress in the field of AI has given rise to a new development: models that can perform a diverse range of tasks across different domains. These models—known as "general purpose AI systems"—exhibit impressive capabilities.[205] For example, DeepMind's Gato model can play video games, caption images, engage in dialogue, and control robotic tools.[206] General purpose systems can sometimes accomplish goals for which they were not explicitly designed.[207] For instance, OpenAI's GPT-3 language model, which was trained to produce English text, learned to write computer code.[208]

General purpose systems can also operate as "foundation models," that is, they can serve as a foundation or platform that underpins many downstream applications.[209] Exploiting the capabilities of these systems in novel contexts is often more efficient than developing new, context-specific systems. General purpose systems and foundation models can typically perform tasks better than narrower systems, or can be adapted to do so at relatively low cost.[210] For example, rather than training a new language model to summarize academic journal articles, it is cheaper and easier to use an existing foundation model. It is therefore no surprise that general purpose systems are among the most popular AI products and have been described as "the future of AI."[211]

---

[205] For a broader discussion of AI as a general purpose technology, see SULEYMAN, *supra* note 33, at ch. 2; Manuel Trajtenberg, *AI as the Next GPT: a Political-Economy Perspective*, (NBER Working Paper No. 24245, Jan. 2018), https://www.nber.org/papers/w24245; Nicholas Crafts, *Artificial Intelligence as a General-Purpose Technology: An Historical Perspective*, 37 OXFORD REV. ECON. POL'Y 521 (2021); Avi Goldfarb, Bledi Taska & Florenta Teodoridis, *Could Machine Learning Be a General Purpose Technology? A Comparison of Emerging Technologies Using Data from Online Job Postings*, 52 RES. POL'Y 104653 (2023); Ben Garfinkel, *The Impact of Artificial Intelligence: A Historical Perspective*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (Justin B. Bullock et al. eds., forthcoming); Tyna Eloundou, Sam Manning, Pamela Mishkin & Daniel Rock, *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, ARXIV (Aug. 21, 2023), https://arxiv.org/abs/2303.10130.

[206] *See* Scott Reed et al., *A Generalist Agent*, 11 TRANSACTIONS MACH. LEARN. RES (Nov. 2022) (introducing the Gato model and describing its capabilities and limitations).

[207] *See* Wei et al., *supra* note 54, at 52.

[208] *See* Brown et al., *supra* note 53 (examining the capabilities of the GPT-3 model).

[209] *See* Bommasani et al., *supra* note 5, at 3; Barrett, Hendrycks, Newman & Nonnecke, *supra* note 180, at 5; Timnit Gebru et al., *Five Considerations to Guide the Regulation of "General Purpose AI" in the EU's AI Act: Policy Guidance from a Group of International AI Experts*, AI NOW INST. at 3 (Apr. 13, 2023), https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act.

[210] Such adaption is typically facilitated through fine-tuning, i.e., partially retraining a large model on a smaller specialized dataset to optimize performance on a particular task.

[211] Kris Shrishak & Risto Uuk, *The EU AI Law Will Not Be Future-Proof Unless It*

The problem with general purpose systems is that any safety risks they pose—including robustness errors, harmful biases, and misalignment with societal values—are likely to propagate downstream.[212] In other words, even minor defects in a general purpose system could have disastrous consequences if the system is deployed at scale or in high-stakes settings, such as healthcare or finance.[213] This risk profile is well known in cybersecurity. For example, in early 2017, a single vulnerability in a widely used software framework led to a security breach in which the financial records of over 150 million consumers were compromised.[214] Similarly, a single safety risk affecting a widely deployed general purpose AI system could cause devastating harm.

Given these weighty concerns, it is unsettling that *none* of the major U.S. regulatory proposals refers to general purpose AI systems.[215] Commentators have pointed out this glaring governance gap, but little action has been taken to remedy it.[216] By contrast, in the European Union, the issue of how to regulate general purpose systems has been the subject of heated debate.[217] While the initial draft of the EU AI Act made no reference to general purpose systems,[218] and was widely criticized for this omission,[219] subsequent drafts of the Act have sought to address the issue directly.[220] These attempts, however, are riddled with difficulties. As will be shown, the regime proposed in the EU AI Act for regulating general purpose systems is far from watertight.

Regulates General Purpose AI Systems, EURACTIV (May 30, 2022), https://www.euractiv.com/section/digital/opinion/the-eu-ai-law-will-not-be-future-proof-unless-it-regulates-general-purpose-ai-systems/. *See also* Gebru et al., *supra* note 209.

[212] *See*, *e.g.*, Shangbin Feng, Chan Young Park, Yuhan Liu & Yulia Tsvetkov, *From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models*, PROC. 61ST ANN. MEETING ASS'N COMPUT. LINGUISTICS 11737, 11743 (2023). In addition, vulnerabilities can be transferable across different AI models. *See* Andy Zou, Zifan Wang, J. Zico Kolter & Matt Fredrikson, *Universal and Transferable Adversarial Attacks on Aligned Language Models*, ARXIV (Jul. 27, 2023), https://arxiv.org/abs/2307.15043.

[213] *See* Bommasani et al., *supra* note 5, at 114–16; Gebru et al., supra *note* 209, at 4.

[214] *See* Gaglione, *supra* note 133, at 1160–66.

[215] The Blumenthal-Hawley Framework refers to general purpose systems. However, no bills have (at the time of writing) been introduced to implement the framework.

[216] *See*, *e.g.*, Barrett, Hendrycks, Newman & Nonnecke, *supra* note 180, at 30.

[217] *See*, *e.g.*, Hodan Omaar, *Should the EU Regulate General-Purpose AI Systems?*, CENTER FOR DATA INNOVATION (Sept. 13, 2022), https://datainnovation.org/2022/09/should-the-eu-regulate-general-purpose-ai-systems/.

[218] EU AI Act (Apr. 21, 2021).

[219] *See*, *e.g.*, Risto Uuk, *General Purpose AI and the AI Act*, FUTURE OF LIFE INST. (May 2022), https://futureoflife.org/wp-content/uploads/2022/08/General-Purpose-AI-and-the-AI-Act-v5.pdf.

[220] EU AI Act (Nov. 25, 2022) rec. 12c.

The EU AI Act defines a general purpose AI system as any AI system that "is intended by the provider to perform generally applicable functions" and "may be used in a plurality of contexts and be integrated in a plurality of other AI systems."[221] Rather than impose stringent requirements on all general purpose systems, the Act provides that only general purpose systems that may be used as high-risk systems are subject to any obligations at all; and those obligations are only a subset of the requirements applicable to other high-risk systems.[222] For example, the obligation to report serious incidents, which applies to other high-risk AI systems, does not apply to general purpose systems.[223]

There is also considerable ambiguity around which entity is responsible for complying with the requirements applicable to general purpose systems under the EU AI Act. While responsibility ordinarily falls on the provider of a general purpose system, the Act stipulates that the provider will no longer be responsible if another party makes a "substantial modification" to the system.[224] The definition of "substantial modification," however, contains several ambiguities.[225] It is therefore unclear in which circumstances responsibility shifts from the upstream developer (who builds a foundation model) to a downstream user (who modifies and deploys that model).

This division of responsibility is particularly concerning if we consider the actors involved in the AI value chain. Developers of general purpose systems are typically well-resourced industry labs, such as OpenAI, Google DeepMind, and Meta, which employ dedicated safety and ethics teams.[226]

---

[221] *Id.* art. 3(1b). *Compare* Carlos Ignacio Gutierrez, Anthony Aguirre, Risto Uuk, Claire Boine & Matija Franklin, *A Proposal for a Definition of General Purpose Artificial Intelligence Systems*, 2 DIGITAL SOC. 36 (2023); Risto Uuk, Carlos Ignacio Gutierrez & Alex Tamkin, *Operationalising the Definition of General Purpose AI Systems: Assessing Four Approaches*, ARXIV (June 5, 2023), https://arxiv.org/abs/2306.02889; Simeon Campos & Romain Laurent, *A Definition of General-Purpose AI Systems: Mitigating Risks from the Most Generally Capable Models* (Working Paper, Jul. 16, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4423706.

[222] EU AI Act arts. 4a(1), 4b(1), 4(b)(6). *See also* arts. 4b(2)–(5) (describing the requirements applicable to general purpose systems used as, or as components of, high-risk systems). *But see* Alex C. Engler & Andrea Renda, *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*, CENTRE FOR EUR. POL'Y STUD. at 20 (Sept. 3, 2022), https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/ ("Given the many categories of AI systems in products and standalone AI systems that can fall into the high-risk category of the AI Act, functionally this means that all [general purpose] systems would trigger these requirements.").

[223] *See* EU AI Act art. 4b(2) (applying certain requirements to general purpose systems, while excluding others, such as the reporting of serious incidents under Article 62).

[224] *Id.* art. 23a(1)–(2).

[225] *Id.* art. 3(23).

[226] *See supra* Part II.A. *See also* Gebru et al., *supra* note 209, at 6; Angela Müller & Natali Helberger, *The AI Act and General Purpose AI*, ALGORITHM WATCH at 4–5 (Sept. 12,

Users of general purpose systems, by contrast, are typically lower-resourced organizations with far less capacity to evaluate, let alone improve, the safety of these systems. By shifting responsibility to these lower-resourced organizations, the Act simultaneously exculpates the actors best placed to mitigate the risks of general purpose systems, and burdens smaller organizations with important duties they lack the resources to fulfil.[227]

Finally, the Act stipulates that a provider of a general purpose AI system is exempt from the requirements applicable to high-risk systems if "the provider explicitly excluded all high-risk uses in the instructions of use or information accompanying the general purpose AI system."[228] In other words, a written manual that politely instructs users to refrain from deploying the system in a sensitive or safety-critical setting will absolve the provider of all responsibility.[229]

The EU AI Act's combination of lax requirements, ambiguity, and sweeping exemptions for general purpose AI systems is troubling. Whether the resulting regime stems from careless oversight, the influence of lobbying,[230] or risk analysis that differs from the consensus within the AI safety community, the outcome is concerning. The world's regulatory "superpower"[231] and first-mover in AI regulation is failing to establish appropriate safeguards around the most consequential AI technology.

B. Proliferation and Misuse

Like many technologies, AI systems are fundamentally dual use tools. While they can be employed in applications that benefit society, they can also be used for malicious purposes.[232] This unavoidable feature of AI could give

---

2023), https://algorithmwatch.org/en/ai-act-general-purpose-ai/.

[227] Arguably, these obligations potentially apply to *no one*, given that small and medium-sized firms are altogether exempt from the requirements applicable to general purpose systems. *See* EU AI Act art. 55a(3).

[228] EU AI Act art. 4c(1).

[229] Gebru et al., *supra* note 209, at 6 ("Any regulatory approach that allows developers of GPAI to relinquish responsibility using a standard legal disclaimer would be misguided.").

[230] As in other domains, lobbying and regulatory capture could influence the content of AI regulation. *See* Bill Perrigo, *OpenAI Lobbied the E.U. to Water Down AI Regulation*, TIME (June 20, 2023), https://time.com/6288245/openai-eu-lobbying-ai-act/. *See also* Clark & Hadfield, *supra* note 120, at 6; Sam Clarke, Jess Whittlestone, Matthijs Maas, Haydn Belfield, José Hernández-Orallo & Seán Ó hÉigeartaigh, *Submission of Feedback to the European Commission's Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence*, UNIVERSITY OF CAMBRIDGE at 5 (Aug. 10, 2021), https://www.cser.ac.uk/resources/feedback-european-regulation/; Crootof & Ard, *supra* note 131, at 382.

[231] BRADFORD, *supra* note 137, at xiii; Bradford, *supra* note 137, at 66–67.

[232] *See* Markus Anderljung & Julian Hazell, *Protecting Society from AI Misuse: When*

rise to black swan events. For example, machine learning models developed to accelerate drug discovery were adapted to design 40,000 chemical warfare agents, in just six hours.[233] Meanwhile, AI systems designed to assist writers and programmers can be repurposed to carry out cyberattacks and perpetrate financial fraud at unprecedented scale.[234]

The dual use risks posed by AI, however, differ from those posed by other technologies. Unlike industrial equipment or potentially hazardous chemicals, AI technologies lend themselves to rapid and widespread diffusion.[235] According to AI safety researchers, "stealing and widely proliferating powerful AI systems could just be a matter of copy and pasting."[236] The number of actors who can deploy AI systems for nefarious purposes is nearly unlimited. Meanwhile, the pace at which these technologies can be adapted to anti-social ends is frighteningly fast.

It is therefore reassuring that several leading AI labs tightly control access to their systems through a combination of technical and legal guardrails.[237]

---

are Restrictions on Capabilities Warranted?, ARXIV (Mar. 29, 2023), https://arxiv.org/abs/2303.09377; Maximilian Mozes, Xuanli He, Bennett Kleinberg & Lewis D. Griffin, *Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities*, ARXIV (Aug. 24, 2023), https://arxiv.org/abs/2308.12833; Hendrycks, Mazeika & Woodside, *supra* note 14, at 6–12; Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV (Feb. 20, 2018), https://arxiv.org/abs/1802.07228. *See also* Anna G. Eshoo (D-CA), *Eshoo Urges NSA & OSTP to Address Unsafe AI Practices*, CONGRESSWOMAN ANNA G. ESHOO (Sept. 22, 2022), https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-add ress-unsafe-ai-practices (explaining that AI systems are "dual-use tools that can lead to real-world harms like the generation of child pornography, misinformation, and disinformation").

[233] Urbina, Lentzos, Invernizzi & Ekins, *supra* note 17, at 189. For broader discussion of biosecurity risks arising from AI tools, see Jonas B. Sandbrink, *Artificial intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools*, ARXIV (Aug. 12, 2023), https://arxiv.org/abs/2306.13952; Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter & Kevin M. Esvelt, *Can Large Language Models Democratize Access to Dual-Use Biotechnology?*, ARXIV (June 6, 2023), https://arxiv.org/abs/2306.03809; Robert F. Service, *Could Chatbots Help Devise the Next Pandemic Virus?*, 380 SCIENCE 1211 (2023).

[234] *See Cybercriminals Starting to Use ChatGPT*, CHECK POINT RES. (Jan. 6, 2023), https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/; Julian Hazell, *Large Language Models Can Be Used to Effectively Scale Spear Phishing Campaigns*, ARXIV (May 12, 2023), https://arxiv.org/abs/2305.06972. *See also* Weidinger et al., *Ethical and Social Risks*, *supra* note 5, at 26–28; Chen et al., *supra* note 80, at 12; Ben Buchanan, John Bansemer, Dakota Cary, Jack Lucas & Micah Musser, *Automating Cyber Attacks: Hype and Reality*, GEORGETOWN CTR. FOR SEC. & EMERGING TECH. (Nov. 2020), https://cset.georgetown.edu/publication/automating-cyber-attacks/.

[235] *See* Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, ARXIV at 13–15 (Sept. 4, 2023), https://arxiv.org/abs/2307.03718. *See also* Shevlane, *Governing Artefacts*, *supra* note 110, at 3; SULEYMAN, *supra* note 33, at chs. 2–3.

[236] Hendrycks & Mazeika, *supra* note 69, at 12.

[237] *See* OpenAI, *Best Practices for Deploying Language Models*, OPENAI (June 2, 2022),

However, this is not the norm. Apart from the fact that relatively few companies adopt measures to reduce proliferation and misuse, on occasion companies have undermined the safety measures taken by their competitors. For example, while OpenAI provides access to its latest models only via API,[238] Meta provides broader access to its models.[239] This is part of a broader trend. A new crop of independent AI labs is rapidly releasing to the public state-of-the-art datasets, code, and training techniques.[240] While this trend toward technological democratization advances scientific and commercial progress in AI,[241] it also heightens the risk of misuse.[242] Malicious actors are now able to utilize a growing collection of publicly available high-quality AI resources that can be repurposed to harmful ends.

In other safety-critical domains, legal mechanisms address proliferation challenges by establishing stringent safeguards. For example, the Federal Aviation Administration imposes onerous requirements on the sale and transfer of commercial aircraft.[243] The Centers for Disease Control and Prevention sets biosafety levels and procedures for dangerous biological

---

https://openai.com/blog/best-practices-for-deploying-language-models/ (outlining a suite of best practices adopted by several leading language model developers).

[238] This began with GPT-3 and has continued with ChatGPT and GPT-4. *See* Greg Brockman, Mira Murati & Peter Welinder, *OpenAI API*, OPENAI (June 11, 2020), https://openai.com/blog/openai-api/; Schulman et al., *supra* note 1; OpenAI, *GPT-4*, OPENAI (Mar. 14, 2023), https://openai.com/research/gpt-4. For an overview of different model release approaches, see Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, PROC. 2023 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 111 (2023).

[239] *See* Joe Spisak & Sergey Edunov, *The Llama Ecosystem: Past, Present, and Future*, META (Sept. 27, 2023), https://ai.meta.com/blog/llama-2-updates-connect-2023/. *But see* Michael Nolan, *Llama and ChatGPT Are Not Open-Source*, IEEE SPECTRUM (Jul. 27, 2023), https://spectrum.ieee.org/open-source-llm-not-open.

[240] *See* Benaich & Hogarth, *supra* note 103, at 84; Benaich, *supra* note 103, at 100.

[241] *See* Alex Engler, *The EU's Attempt to Regulate Open-Source AI is Counterproductive*, BROOKINGS INST. (Aug. 24, 2022), https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/. *See also* Kyle Wiggers, *The EU's AI Act Could Have a Chilling Effect on Open Source Efforts*, TECHCRUNCH (Sept. 6, 2022), https://techcrunch.com/2022/09/06/the-eus-ai-act-could-have-a-chilling-effect-on-open-source-efforts-experts-warn/; Sharon Goldman, *Hugging Face, GitHub and More Unite to Defend Open Source in EU AI Legislation*, VENTUREBEAT (Jul. 26, 2023), https://venturebeat.com/ai/hugging-face-github-and-more-unite-to-defend-open-source-in-eu-ai-legislation/.

[242] *See* Elizabeth Seger et al., *Open-Sourcing Highly Capable Foundation Models*, CENTRE FOR THE GOVERNANCE OF AI at 12–16 (Sept. 29, 2023), https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

[243] *See Aircraft Registration*, FED. AVIATION ADMIN., https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/. For discussion of licensing of AI models, see Anderljung et al., *supra* note 235, at 20–21.

research such as pathogen synthesis.[244] The Nuclear Regulatory Commission controls the proliferation of nuclear materials and technology through rigorous licensing, oversight, and enforcement.[245]

Given the prospect of malicious actors using AI technologies to cause large-scale harm, we might expect policymakers to adopt safeguards comparable to those used in aviation, biosafety, or nuclear energy.[246] But the main current proposals for regulating AI disappoint. The principles enshrined in the White House's AI Bill of Rights and the practices outlined in NIST's AI Risk Management Framework make scant reference to misuse and proliferation.[247] Whether federal regulations or agency actions will fill this gap in the future is an open question.

The EU AI Act is hardly better. Despite enumerating a long list of policy objectives,[248] the Act does not include mitigating misuse among them. The risk of malicious use appears only twice in the Act's eighty-five operative provisions.[249] Perhaps most unsettling is the carveout for AI systems developed for "scientific research and development," which are excluded from all requirements under the EU AI Act.[250] This sweeping exemption makes little sense when we consider the risk of misuse. A lack of commercial application has no bearing on the ability of malicious actors to employ an AI system toward harmful ends. As illustrated above, even systems developed

---

[244] *See* BIOSAFETY IN MICROBIOLOGICAL AND BIOMEDICAL LABORATORIES (Centers for Disease Control and Prevention, 6th ed. 2020), https://www.cdc.gov/labs/BMBL.html. For a comparable framework in AI development, see *Anthropic's Responsible Scaling Policy*, ANTHROPIC (Sept. 19, 2023), https://www.anthropic.com/index/anthropics-responsible-scaling-policy ("defines a framework called AI Safety Levels (ASL) for addressing catastrophic risks, modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials.").

[245] *See Licensing of Medical, Industrial, and Academic Uses of Nuclear Materials*, NUCLEAR REGULATORY COMMISSION, https://www.nrc.gov/materials/miau/licensing.html. For discussion of licensing for AI models, see Anderljung et al., *supra* note 235, at 20–21.

[246] *See* Eshoo, *supra* note 232 ("In the same way that nuclear information and materials may lead to both the generation of energy and horrible atrocities, AI models similarly pose dual-use applications … We currently use export controls to control the release of various types of dual-use technical data, and I urge you to investigate the possibility of using such powers to control the release of unsafe dual-use AI models as well.").

[247] *See* AI Bill of Rights, *supra* note 182, at 18; NIST AI RMF, *supra* note 174, at 15.

[248] EU AI Act art. 1.

[249] *Id.* art. 14(2) (requiring human oversight of certain AI systems in order to, *inter alia*, prevent or minimize "reasonably foreseeable misuse"). This is the only operative provision in the Act to mention misuse, other than in the context of general purpose systems. *See id.* art. 4c(3) (requiring providers of general purpose systems, upon detecting or being informed of misuse, to "take all necessary and proportionate measures to prevent such further misuse.") In addition to the term "proportionate" significantly relaxing the obligations of providers, the requirement itself arguably arrives too late in the AI lifecycle, namely only *after* misuse has already occurred.

[250] *Id.* arts. 2(6)–(7).

solely for (prosocial) research purposes, such as drug discovery, can be weaponized.

That policymakers have overlooked these concerns, or dismissed them, is worrying. Given the rapid pace at which state-of-the-art AI systems are becoming publicly available, the threat of malicious actors using these systems to commit crime and cause harm is pervasive and imminent. While other safety-critical domains offer important insights and concrete approaches for mitigating misuse risks, policymakers in the field of AI have not, as yet, heeded the lesson.

C. Systemic Risk

Apart from the risk of large-scale harm caused by misuse or malfunction, the widespread deployment of powerful AI systems presents another type of risk: the gradual erosion of social and political institutions and values.[251] Systemic risks of this kind, to an even greater extent than other algorithmic black swans, are a complex sociotechnical phenomenon.[252] The incentives of AI developers, the design of AI systems, and the way users interact with these systems can, together, undermine vital components of a well-functioning democracy. Consider the following example, familiar to us from YouTube, Netflix, and TikTok: *content selection algorithms*.

> [These algorithms] aren't particularly intelligent, but they are in a position to affect the entire world because they directly influence billions of people. Typically, such algorithms are designed to maximize *click-through*, that is, the probability that the user clicks on presented items. The solution is simply to present items that the user likes to click on, right? Wrong. The solution is to change the user's preferences so that they become more predictable. A more predictable user can be fed items that they are likely to click on, thereby generating more revenue. People with more extreme political views tend to be more predictable in which items they will click on.[253]

---

[251] *See* Nathalie A. Smuha, *Beyond the Individual: Governing AI's Societal Harm*, 10 INTERNET POL'Y REV. 1 (2021); Remco Zwetsloot & Allan Dafoe, *Thinking About Risks From AI: Accidents, Misuse and Structure*, LAWFARE (Feb. 11, 2019), https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

[252] *See generally* Steven L. Schwarcz, *Systemic Risk*, 97 GEO. L.J. 193, 198–204 (2008); Kevin Werbach & David Zaring, *Systemically Important Technology*, 101 TEX. L. REV. 811, 832–42 (2023); Yacov Y. Haimes, *On the Complex Definition of Risk: A Systems-Based Approach*, 29 RISK ANALYSIS 1647 (2009); Ortwin Renn, Manfred Laubichler, Klaus Lucas, Wolfgang Kröger, Jochen Schanze, Roland W. Scholz & Pia-Johanna Schweizer, *Systemic Risks from Different Perspectives*, 42 RISK ANALYSIS 1902 (2022) (discussing the major properties of systemic risk, namely complexity, ambiguity, and ripple-effects).

[253] RUSSELL, *supra* note 14, at 8. *See also* Gilbert, Dean, Zick & Lambert, *supra* note 75, at 29 ("By slotting the user into behaviors whose engagement is easier to control, rather than showing whatever content happens to have worked on others, [reinforcement learning]

This disturbing observation, unfortunately, is not a speculative prediction. For over a decade we have known that AI systems can manipulate people's preferences.[254] More recently, studies have found that AI systems deployed by social media companies have promoted politically divisive content,[255] incited physical violence,[256] and influenced citizens' voting behavior.[257] While the harms to individuals or groups of people are significant, the aggregate harm to social and political values is even more dramatic.[258]

---

could "hack" users to socialize through the platform, rather than other forms of civic participation."). Researchers are now actively developing systems with these properties. *See, e.g.*, Xueliang Wang, *Reinforcing User Retention in a Billion Scale Short Video Recommender System*, COMPANION 2023 PROC. ACM WEB CONF. 421 (2023); Robert Irvine et al., *Rewarding Chatbots for Real-World Engagement with Millions of Users*, ARXIV (Mar. 30, 2023), https://arxiv.org/abs/2303.06135. For a systematic treatment of these issues, see Jonathan Stray et al., *Building Human Values into Recommender Systems: An Interdisciplinary Synthesis*, ARXIV (Jul. 20, 2022), https://arxiv.org/abs/2207.10192.

[254] *See, e.g.*, Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley & Jingjing Zhang, *Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects*, 24 INFO. SYS. RES. 883 (2013). For recent studies, see Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson & Mor Naaman, *Co-Writing with Opinionated Language Models Affects Users' Views*, PROC. 2023 CHI CONF. ON HUMAN FACTORS IN COMPUT. SYS. (2023); Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao & Anca D. Dragan, *Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media*, ARXIV (Sept. 14, 2023), https://arxiv.org/abs/2305.16941; Giovanni Spitale, Nikola Biller-Andorno & Federico Germani, *AI Model GPT-3 (Dis)informs Us Better than Humans*, 9 SCIENCE ADVANCES eadh1850 (2023); Celeste Kidd & Abeba Birhane, *How AI Can Distort Human Beliefs*, 380 SCIENCE 1222 (2023).

[255] *See* Jack Nicas, *How YouTube Drives People to the Internet's Darkest Corners*, WALL. ST. J. (Feb. 7, 2018), https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.

[256] *See* Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. TIMES (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

[257] *See* Robert M. Bond et al., *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, 589 NATURE 295 (2012) (showing that social media messaging can influence political self-expression, information seeking, and voting behavior). *See also* Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky & Ralph Hertwig, *A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy*, 7 NATURE HUMAN BEHAV. 74 (2023). *Compare* Mike Isaac & Sheera Frenkel, *Facebook's Algorithm Is 'Influential' but Doesn't Necessarily Change Beliefs, Researchers Say*, N.Y. TIMES (Aug. 3, 2023) (surveying recent studies finding that social media has a less significant political impact).

[258] *See* Kaminski, *supra* note 36, at 146 ("Assessing risk also typically means discussing harms at the level of the collective. That is, rather than preventing or compensating for individualized harms, risk thinking assesses harms at a social level. It aims at the bigger picture, at populations and systems rather than at persons"). *See also* Edwards, *supra* note 150, at 2 ("Impacts on groups and on society as a whole need to be considered, as well as risks to individuals and their rights").

Polarization and radicalization can reduce trust in democratic institutions and cause profound social disruption.[259] Even subtle changes in the beliefs of a small number of people can impact election outcomes, which can in turn have far-reaching societal implications.[260]

There is little indication that these risks will abate. The business model of Google, Meta, and other leading AI developers has not changed. The underlying technology—AI systems that optimize simple reward functions (such as naively satisfying user preferences)—remains popular. If anything, the magnitude and frequency of systemic risks will increase as AI systems become more capable and are used more widely.[261] For example, while content selection algorithms can *recommend* polarizing content to users, AI systems that generate text, images, and video can *create* polarizing content that targets the specific characteristics and vulnerabilities of individual users.[262]

---

[259] *See, e.g.*, Jack Citrin & Laura Stoker, *Political Trust in a Cynical Age*, 21 ANN. REV. POL. SCI. 49, 59 (2018) (discussing the impact of polarization on political trust).

[260] *See* Future of Life Institute, *Response to Request for Information: AI RMF* (Sept. 15, 2021), https://www.nist.gov/document/ai-rmf-rfi-comments-future-life-institute ("Though the vast majority of users may be relatively unaffected by being unintentionally recommended polarizing or radicalizing content, if even a small percentage (e.g., 0.1%) does evince negative effects, it can create a societal-wide consequence."); Future of Life Institute, *FLI Position Paper on the EU AI Act* at 4 (Aug. 4, 2021), https://futureoflife.org/wp-content/uploads/2021/08/FLI-Position-Paper-on-the-EU-AI-Act.pdf?x76795 ("AI applications may cause societal-level harms, even when they cause only negligible harms to individuals. For example, a political marketing application may reduce a person's desire to vote by a small amount. At an individual level, the impact of this application may not be considered an infringement of fundamental rights, but collectively, the effect may be large enough to change an election result"). *See also* Smuha, *supra* note 251, at 10 ("societal harm typically does not arise from a single occurrence of the problematic AI practice. Instead, it is often the widespread, repetitive or accumulative character of the practice that can render it harmful from a societal perspective.").

[261] *See* Michael Guihot, Anne F. Matthew & Nicolas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 416 (2017) (explaining that the integration of AI into "complex, interdependent social, financial, and economic systems or networks amplifies the potential for risk … The more complex and nonlinear these networks are, the easier it is for the impacts of an Al incident to proliferate rapidly throughout the network, affecting multiple stakeholders.").

[262] *See* Matthew Burtell & Thomas Woodside, *Artificial Influence: An Analysis of AI-Driven Persuasion*, ARXIV (Mar. 15, 2023), https://arxiv.org/abs/2303.08721; Hui Bai, Jan Voelkel, Johannes Eichstaedt & Robb Willer, *Artificial Intelligence Can Persuade Humans on Political Issues* (Sept. 7, 2023), https://www.researchsquare.com/article/rs-3238396/v1. *See also* Ben Buchanan, Andrew Lohn, Micah Musser & Katerina Sedova, *Truth, Lies, and Automation How Language Models Could Change Disinformation*, GEORGETOWN CTR. FOR SEC. & EMERGING TECH. (May 2021), https://cset.georgetown.edu/publication/ truth-lies-and-automation/; Weidinger et al., *Ethical and Social Risks*, *supra* note 5, at 25–26; Bommasani et al., *supra* note 5, at 135–38 (explaining that foundation models facilitate creating personalized content at low cost); Micah Musser, *A Cost Analysis of Generative*

Some of these systemic risks are likely to result from deliberate misuse. For example, malicious actors could use generative AI systems to conduct large-scale misinformation campaigns or flood lawmakers with high-quality automated comments and requests, distorting their perceptions of the public interest and endangering vital political processes.[263] Other systemic risks, however, can arise inadvertently, resulting from defective or misaligned systems. For instance, recent research illustrates that more powerful language models tend to express stronger political views, including on gun rights and immigration, and are more "sycophantic," that is, they "are more likely [than less powerful models] to answer questions in ways that create echo chambers by repeating back a … user's preferred answer."[264]

Equally concerning is the prospect of AI systems polluting or systematically manipulating our information environment. Just as the widespread use of ChatGPT could degrade the quality of answers in the Stack Overflow programming forum,[265] other generative systems could cause tremendous harm to information utilities such as Wikipedia and YouTube.[266] Gary Marcus, a prominent AI researcher, argues that the combination of these systems' unreliability and their ability to be cheaply deployed at scale "pose a real and imminent threat to the fabric of society."[267]

Language Models and Influence Operations, ARXIV (Aug. 7, 2023), https://arxiv.org/abs/2308.03740. *See generally* Ben M. Tappin et al., *Quantifying the Potential Persuasive Returns to Political Microtargeting*, 120 PROC. NAT'L ACAD. SCI. e2216261120 (2023).

[263] *See* Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel & Katerina Sedova, *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, ARXIV (Jan. 10, 2023), https://arxiv.org/abs/2301.04246; Josh A. Goldstein & Girish Sastry, *The Coming Age of AI-Powered Propaganda*, FOREIGN AFFAIRS (Apr. 7, 2023). *See also* John J. Nay, *Large Language Models as Corporate Lobbyists*, ARXIV (Jan. 5, 2023), https://arxiv.org/abs/2301.01181; Tyler Cowen, *ChatGPT AI Could Make Democracy Even More Messy*, BLOOMBERG (Dec. 6, 2022), https://www.bloomberg.com/opinion/articles/2022-12-06/chatgpt-ai-could-make-democracy-even-more-messy.

[264] Perez et al., *Discovering Language Model Behaviors*, *supra* note 15, at 13388. *See also* Mina Lee et al., *Evaluating Human-Language Model Interaction*, ARXIV at 17 (Dec. 19, 2022), https://arxiv.org/abs/2212.09746 (explaining that language models may "influence human writing practices, opinions and beliefs … and potentially many other aspects of human experience that are mediated by language.") *See also* Bommasani et al., *supra* note 5, at 130 (noting that foundation models "encode an Anglocentric perspective by default, which can amplify majority voices and contribute to homogenization of perspectives or monoculture"). *See also id.* at 151–52; Jon Kleinberg & Manish Raghavan, *Algorithmic Monoculture and Social Welfare*, 118 PROC. NAT'L ACAD. SCI. e2018340118 (2021).

[265] *See* Maria del Rio-Chanona, Nadzeya Laurentsyeva & Johannes Wachs, *Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow*, ARXIV (Jul. 14, 2023), https://arxiv.org/abs/2307.07367.

[266] *Id.*

[267] Gary Marcus, *AI's Jurassic Park Moment*, THE ROAD TO AI WE CAN TRUST (Dec.

To date, the response of policymakers has been underwhelming. In the United States, the White House AI Bill of Rights is mainly couched in terms of risks to *individual* rights, rather than risks to social and political institutions. All five principles it enshrines appear in the second-person. For example, "*You* should be protected from abusive data practices" and "*you* should have agency over how data about you is used."[268] While such individual protections are necessary, they are not sufficient. The NIST AI Risk Management Framework pays lip service to potential systemic risks, referring to both the "individual and societal impacts related to AI risks."[269] But the framework does not elaborate on the nature of these large-scale risks or how it plans to tackle them.[270]

The European Union's response is somewhat more encouraging.[271] In determining whether an AI system presents a high risk, the EU AI Act takes into account the anticipated impact on "society at large."[272] In addition, the Act explicitly prohibits AI systems that are intended or likely to manipulate human behavior.[273] However, this prohibition arguably focuses more on

---

10, 2022), https://garymarcus.substack.com/p/ais-jurassic-park-moment.

[268] White House AI Bill of Rights, *supra* note 182, at 6.

[269] NIST AI Risk Management Framework, *supra* note 174, at 24. *See also id.* at 2, 40. *See also* Future of Life Institute, *AI RMF Comments* (Sept. 15, 2021), https://www.nist.gov/system/files/documents/2021/09/17/ai-rmf-rfi-0106-attachment.pdf (suggesting that NIST should "consider is the aggregate systematic impact of small effects by AI systems that, when deployed on a massive scale, can lead to harms of a societal magnitude"); Jonas Schuett & Markus Anderljung, *Submission to the NIST AI Risk Management Framework*, CENTRE FOR THE GOVERNANCE OF AI at 2 (May 4, 2022), https://www.governance.ai/research-paper/submission-to-the-nist-ai-risk-management-framework (arguing that NIST's proposal does not adequately address large-scale risks to society).

[270] A similar critique could be directed toward the National AI Commission Act § 3(g)(1) (tasking the Commission with "establishing necessary, long-term guardrails to ensure that artificial intelligence is aligned with values shared by all Americans" but giving little indication as to how such guardrails could be constructed).

[271] While the focus of this article is on regulatory efforts targeting AI, other EU regulations may be relevant. *See, e.g.*, Digital Services Act, 2022 O.J. (L 277) 1 [hereinafter DSA] art. 34–35 (requiring that "very large online platforms … diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems or from the use made of their services" and "put in place reasonable, proportionate and effective mitigation measures, tailored to … specific systemic risks …"). The application (and effectiveness) of the DSA vis-à-vis systemic risks from AI will turn on, *inter alia*, whether AI developers and providers are designated as "very large online platforms" pursuant to the procedure stipulated in Article 33. For broader discussion of the DSA, see Ioanna Tourkochoriti, *The Digital Services Act and the EU as the Global Regulator of the Internet*, 24 CHI. J. INT'L L. 129 (2023).

[272] EU AI Act art. 7(2)(i).

[273] *Id.* art. 5(1) (the prohibition applies to "an AI system that deploys subliminal techniques beyond a person's consciousness with the objective to or the effect of materially distorting a person's behaviour in a manner that causes or is reasonably likely to cause that

psychological harm caused to an individual than on the societal implications of large-scale manipulation.[274] Moreover, the few mechanisms in the EU AI Act that aim to mitigate systemic risk, including the monitoring and reporting of safety incidents, are not particularly robust.[275] It is also unclear whether the European Union has the institutional capacity to effectively implement these protective measures.

The common theme behind these shortcomings of the EU AI Act is its focus on product safety, which is a fundamentally individualistic regulatory paradigm.[276] Rather than address the broader, longer-term implications of unsafe AI technologies, the Act primarily targets the immediate risks to individual consumers.[277] Considering the noteworthy risks AI poses to social and political institutions, this individual-centric regulatory approach is inappropriate. But unfortunately, by virtue of the Brussels Effect, the approach is already diffusing globally. For instance, Canada's draft Artificial Intelligence and Data Act proposes restrictions on AI systems that "may result in serious harm to *individuals* or harm to their interests."[278] The Act defines harms as "physical or psychological harm to an *individual*, damage to an *individual's* property, or economic loss to an *individual*."[279] Although

---

person or another person physical or psychological harm").

[274] *See* Veale & Zuiderveen Borgesius, *supra* note 146, at 100 ("Manipulative AI systems appear permitted insofar as they are unlikely to cause an individual (not a collective) 'harm'.") *See also* Matija Franklin, Hal Ashton, Rebecca Gorman & Stuart Armstrong, *Missing Mechanisms of Manipulation in the EU AI Act*, 35 PROC. FLA. AI RES. SOC. (2022); Risto Uuk, *Manipulation and the AI Act*, FUTURE OF LIFE INST. (Jan. 18, 2022), https://futureoflife.org/wp-content/uploads/2022/08/FLI-Manipulation_AI_Act.pdf. Similar issues arise in the EU AI Liability Directive. *See* Prettner, *supra* note 161, at 6–7 (explaining that it is "not clear whether the Directive would allow for broader societal harms caused by AI systems to be covered, such as manipulation at scale, election interference or environmental harms.").

[275] For example, the regime for reporting serious incidents set out in Article 62 relies on self-reporting, and there is no apparent mechanism for oversight or enforcement. In addition, the regime does not apply to general purpose systems. *See supra* note 223.

[276] *See* Veale & Zuiderveen Borgesius, *supra* note 146, at 98–112.

[277] *See* Edwards, *supra* note 150, at 11; UC Berkeley Center for Human-Compatible AI, *Position Paper on the EU AI Act*, at 3–4 (Aug. 6, 2021), https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665648_en.

[278] Canada AI and Data Act art. 4(b). *But see The Artificial Intelligence and Data Act (AIDA) – Companion Document* (Mar. 13, 2023), https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document (referring to "both impacts on individuals and potential systemic impacts", as well as "[s]ystems that can influence human behaviour at scale" and "collective harms").

[279] *Id*. art. 5(1). *See also* Jamie Duncan & Wendy H. Wong, *Data Rights Will Not Save Democracy*, SCHWARTZ REISMAN INST. FOR TECH. & SOC. (Nov. 22, 2022), https://srinstitute.utoronto.ca/news/data-rights-will-not-save-democracy (arguing that the Canada AI and Data Act fails to address societal-level risks to communities and democratic

individuals certainly deserve robust protection from unsafe AI systems, the risks to social and political institutions should not be overlooked. They too must be addressed.

V. ALGORITHMIC PREPAREDNESS

Regulatory design is difficult at the best of times. It is especially difficult when we are concerned about large-scale risks arising from complex sociotechnical phenomena. Clearly, there is no catch-all solution for addressing the prospect of algorithmic black swans. But that does not relieve policymakers of their responsibility to take steps to mitigate the societal risks posed by AI systems. Ambitious regulatory objectives of this kind are commonplace in other high-stakes settings, ranging from public health and climate policy to cybersecurity and nuclear energy.[280] Given that AI is still maturing as a field, this Article does not offer hard and fast rules for governing the technology and its applications.[281] Instead, it proposes a roadmap for "algorithmic preparedness"—a set of five forward-looking principles to guide the development of regulations that confront the risk of algorithmic black swans and mitigate the harms they pose to society.

> Principle 1: AI governance should aim to anticipate and mitigate large-scale societal harm from AI systems.
>
> Principle 2: AI governance should adopt a portfolio approach comprised of diverse and uncorrelated regulatory strategies.
>
> Principle 3: AI governance should be highly scalable.
>
> Principle 4: AI governance should continually explore and evaluate new regulatory strategies.
>
> Principle 5: Cost-benefit analysis of AI governance interventions should place greater weight on worst-case outcomes.

Before exploring each of these principles in detail, it is important to clarify that algorithmic preparedness is not a comprehensive regulatory framework or a definitive playbook for policymakers. Rather, the objective of this set of guiding principles is to highlight several institutional features that are key to tackling algorithmic black swans and are currently neglected

---

institutions, such as election manipulation and misinformation campaigns).

[280] *See* Hendrycks, Mazeika & Woodside, *supra* note 14, at 12, 33.

[281] *See* Weissinger, *supra* note 94, at 8 (arguing that the goals of AI governance are difficult to specify precisely, compared with the goals of traditional safety regulation such as aviation safety).

by policymakers. The hope is that these principles will lay the groundwork for developing concrete mechanisms that fill the salient governance gaps and assist regulators in confronting the transformative impact of AI technologies.

A. Anticipation

> Principle 1: AI governance should aim to anticipate and mitigate large-scale societal harm from AI systems.

This first principle concerns the goals of AI governance. It suggests that AI governance should, among other things, prioritize tackling algorithmic black swans and that, doing so, requires anticipating those risks in advance. Proactively preventing and mitigating large-scale risks is important for several reasons. First, early intervention can be significantly cheaper than efforts to remedy harms after the fact.[282] Second, preventative measures can protect people and institutions in a way that preserves the option of making different regulatory choices in the future (while acting otherwise forecloses this option).[283] Third, certain harms from AI may be irreversible or catastrophic to the extent that ex post actions, such as compensation, cannot effectively undo or redress the harm caused.[284]

---

[282] *See* Kaminski, *supra* note 36, at 121. For a recent proposal of ex ante AI governance, see Gianclaudio Malgieri & Frank A. Pasquale, *From Transparency to Justification: Toward Ex Ante Accountability for AI* (Brooklyn Law School Legal Studies Paper No. 712, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099657. Others, however, suggest that attempts to address such risks in advance are futile. *See* Choi, *supra* note 65, at 44 ("as long as software errors remain inevitable, the software liability paradigm must shift from prevention to mitigation.") *See id.* at 46 ("Bugs and vulnerabilities are so rampant across the industry that the question of cybercrashes and cyberattacks is not 'whether' but 'when.'").

[283] *See* Cass R. Sunstein, *Irreversible and Catastrophic*, 91 CORNELL L. REV. 841, 857 (2006) ("when regulators are dealing with an irreversible loss, and when they are uncertain about the timing and likelihood of that loss, they should be willing to pay a sum—the option value—in order to maintain flexibility for the future."); POSNER, *supra* note 97, at 161–63 (applying similar arguments in the context of climate regulation); Crootof & Ard, *supra* note 131, at 385 (observing that a preventative approach "facilitates information gathering while extending the time frame during which it is possible to craft effective regulation"). *See also* DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 11 (1980) ("If a technology can be known to have unwanted social effects only when these effects are actually felt, what is needed is some way of retaining the ability to exercise control over a technology even though it may be well developed and extensively used.") *See id.* at 12 ("Since the future is extremely uncertain, options which allow the decision maker to respond to whatever the future brings are to be favoured. Decisions, in other words, ought to be reversible, corrigible, and flexible.").

[284] *Cf.* Gary E. Marchant & Yvonne A. Stevens, *Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologie*s, 51 U.C. DAVIS L. REV. 233, 236 (2017) ("Instead of hopelessly trying to anticipate and quantify unascertainable future risks associated with emerging technologies, resilience relies on a trial-and-error approach that

Consider the following situation. Malicious actors discover a vulnerability in AI systems that control safety-critical aviation infrastructure. These actors exploit the vulnerability to orchestrate an attack on the scale of 9/11, or greater, which in turn triggers devastating geopolitical conflict. No monetary sum can compensate for the losses suffered. The only appropriate course of action is to attempt to prevent the occurrence of such a calamity in the first place.

Anticipating and mitigating algorithmic black swans like this face several challenges. The first is epistemic. As in many regulatory contexts, policymakers have only limited information about the risks they aim to address.[285] The staggering pace of AI development, which routinely surprises even industry insiders,[286] makes the problem particularly acute.[287] Moreover, there is little consensus on what interventions would successfully mitigate large-scale societal risks from AI.[288]

The conventional response to this challenge, outlined in the 2011 White House Memorandum on Principles for Regulation and Oversight of Emerging Technologies, is to develop regulations with "sufficient flexibility to accommodate new evidence and learning and to take into account the evolving nature of information related to emerging technologies and their applications."[289] Implementing this guidance is not straightforward.

---

seeks to aggressively explore the potential benefits of a new technology while remaining vigilant and ready to respond to any emerging harms"). *Compare* Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715, 1734 (2015) (arguing that the informational advantages gained by postponing regulatory intervention are more limited than generally assumed).

[285] For discussion of related issues in environmental regulation, see Bradley C. Karkkainen, *Bottlenecks and Baselines: Tackling Information Deficits in Environmental Regulation*, 86 TEX. L. REV. 1409 (2008); Bradley C. Karkkainen, *Information as Environmental Regulation: TRI and Performance Benchmarking, Precursor to a New Paradigm?*, 89 GEO. L.J. 257 (2001).

[286] *See supra* Part I.B.

[287] *See* Kaminski, *supra* note 36, at 153 ("it is exceedingly hard if not impossible to know, measure, and mitigate all risks in advance. This is especially true where there are unknown unknowns, including potentially catastrophic risk.") *See id.* at 155 ("a … central problem of AI risk regulation is that the risks raised by AI systems are varied, not always quantifiable, often contested, and sometimes excruciatingly or even impossibly hard to define.") *See also* Matthew T. Wansley, *Regulation of Emerging Risks*, 69 VAND. L. REV. 401, 403 (2016) ("Emerging risks differ from other risks that the state regulates … the information necessary to answer potentially dispositive questions about how the risk should be regulated will not be available when regulators first become aware of the technology.").

[288] Over eight hundred different AI governance regimes have been proposed globally. *See National AI Policies & Strategies*, OECD AI POLICY OBSERVATORY, https://oecd.ai/en/dashboards/overview.

[289] John P. Holdren, Cass R. Sunstein & Islam A. Siddiqui, *Memorandum for the Heads of Executive Departments and Agencies on Principles for Regulation and Oversight of Emerging Technologies* at 2 (Mar. 11, 2011), https://obamawhitehouse.archives.gov/sites/

Regulatory learning is notoriously difficult, especially when it requires keeping pace with technological change.[290] David Collingridge famously described the dilemma as follows:

> [T]he social consequences of a technology cannot be predicted early in the life of the technology. By the time undesirable consequences are discovered, however, the technology is often so much part of the whole economic and social fabric that its control is extremely difficult. . . . When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult and time consuming.[291]

The Collingridge dilemma, as the problem has come to be known, is an enduring challenge for regulators. In the case of AI, the imbalance between the resources invested in developing new technological systems compared with the resources invested in governing those systems is large and growing.[292] By default, regulatory action is slow and inflexible,[293] which is clearly inappropriate for addressing the impact of a technology characterized

---

default/files/omb/inforeg/for-agencies/Principles-for-Regulation-and-Oversight-of-Emerging-Technologies-new.pdf.

[290] *See* Guihot, Matthew & Suzor, *supra* note 261, at 421 ("This pacing problem plagues the regulation of technology generally and often leads to the technology disengaging or decoupling from the regulation that seeks to regulate it. Because Al is at the forefront of scientific discovery and is developing so quickly, it is affected by this issue more than other technologies."). For further discussion of the pacing problem, see WENDELL WALLACH, A DANGEROUS MASTER: HOW TO KEEP TECHNOLOGY FROM SLIPPING BEYOND OUR CONTROL 395 (2015); Gary E. Marchant, *Addressing the Pacing Problem*, *in* THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT 199 (Gary E. Marchant, Braden R. Allenby & Joseph R. Herkert eds., 2011). For a critique of the term "pacing problem," see Ryan Calo, *The Scale and the Reactor* at 22 (Apr. 15, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id= 4079851.

[291] COLLINGRIDGE, *supra* note 283, at 11. *See also* Crootof & Ard, *supra* note 131, at 381 ("The efficacy of any legal response to technologically created uncertainty is partially a product of its timing … delay may blunt the impact of the regulation or even render regulation impossible, if the technological design or use has already stabilized or if significant users have invested in the infrastructure. The more time passes, the more likely it is that the technology's design or uses will become entrenched and therefore resistant to regulation.").

[292] *See* Clark & Hadfield *supra* note 120, at 6; Kaminski, *supra* note 36, at 122; Scherer, *supra* note 125, at 387. *See also* Jess Whittlestone & Jack Clark, *Why and How Governments Should Monitor AI Development*, ARXIV at 3 (Aug. 31, 2021), https://arxiv.org/abs/2108.12427 ("The result is a situation where companies are able to deploy AI systems with substantial potential for harm or misuse in mostly unregulated markets, governments are caught off-guard by these new applications and their impacts, and are unable to effectively scrutinize systems in the ways needed to govern them.").

[293] *See* Wansley, *supra* note 287, at 404 (arguing that the requirements of the conventional rulemaking process present "insurmountable obstacles to regulating emerging risks").

by the mantra "move fast and break things."[294] The EU AI Act offers a concrete example of the problem. The Act's mechanism for updating which uses of AI are considered high-risk and trigger more stringent compliance requirements involves a cumbersome and brittle administrative process.[295]

What can policymakers do differently? How can they overcome the Collingridge dilemma and design more adaptive, forward-looking AI regulations?[296] The first step is to equip policymakers with up-to-date and accurate information about the capabilities and impact of AI technologies.[297] This information can be collected in several ways. Policymakers could, either themselves or through third party contractors, monitor and measure the capabilities of state-of-the-art AI systems.[298] Alternatively, policymakers

---

[294] *See* JONATHAN TAPLIN, MOVE FAST AND BREAK THINGS: HOW FACEBOOK, GOOGLE, AND AMAZON CORNERED CULTURE AND UNDERMINED DEMOCRACY 9 (2017) (quoting Mark Zuckerberg: "Move fast and break things. Unless you are breaking stuff, you aren't moving fast enough.") The phrase served as Facebook's internal motto until 2014.

[295] *See* EU AI Act arts. 7(1)–(2). *See also* Clarke, Whittlestone, Maas, Belfield, Hernández-Orallo & Ó hÉigeartaigh, *supra* note 230, at 4 ("Even where provisions for adaptability exist in principle, historical experience suggests that updating regulations frequently or quickly enough can be challenging. For instance, various arms control regimes have struggled to update control lists in a frequent and timely fashion.").

[296] Notable attempts to facilitate adaptive governance include J. B. Ruhl, *General Design Principles for Resilience and Adaptive Capacity in Legal Systems – With Applications to Climate Change Adaptation*, 89 N.C. L. REV. 1373 (2011); Robin Kundis Craig, "*Stationarity is Dead" – Long Live Transformation: Five Principles for Climate Change Adaptation Law*, 34 HARV. ENVTL. L. REV. 9 (2010).

[297] *See* Holdren, Sunstein & Siddiqui, *supra* note 289, at 1–2 ("Federal regulation and oversight of emerging technologies should be based on the best available scientific evidence. Adequate information should be sought and developed, and new knowledge should be taken into account when it becomes available.").

[298] *See* Jack Clark, *Information Markets and AI Development*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE (Justin B. Bullock et al. eds., forthcoming). *See also* Whittlestone & Clark, *supra* note 292, at 1; *id.* at 4 (noting that "governments use information and metrics to manage and oversee many critical policy areas. For example, metrics like inflation are critical for managing the economy, data about the prevalence of traffic on major roadways is an input into infrastructure planning, and during COVID-19 we've seen how basic data about the medical status of citizens is a fundamental input into policymaking."). This has been proposed in some recent legislative frameworks. *See, e.g.*, Blumenthal-Hawley Framework (proposing that audits be conducted by an independent oversight body). For further discussion of auditing mechanisms, see Jakob Mökander, Jonas Schuett, Hannah Rose Kirk & Luciano Floridi, *Auditing Large Language Models: A Three-Layered Approach*, AI & ETHICS (2023); Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg & Daniel Ho, *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, PROC. 2022 AAAI /ACM CONF. AI, ETHICS & SOC'Y 557 (2022); Gregory Falco et al., *Governing AI Safety Through Independent Audits*, 3 NATURE MACH. INTELL. 566 (2021); Toby Shevlane et al., *Model Evaluation for Extreme Risks*, ARXIV (Sept. 22, 2023), https://arxiv.org/abs/2305.15324. For discussion of the broader regulatory context, see Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119

could establish mandatory disclosure regimes whereby AI developers audit their own systems and report their findings to policymakers.[299] Finally, policymakers could, through a combination of whistleblower protection and carefully crafted financial incentives, empower individuals with access to relevant information (such as software engineers at leading AI labs) to disclose information pertaining to automated systems that pose large-scale societal risks.[300]

Equipped with these insights, policymakers will be able to better identify new risks posed by AI technologies and make more informed decisions on how to address them. For example, European regulators could apply their knowledge of the latest technical developments to periodically review which applications are deemed high-risk under the EU AI Act. Similarly, regulators in the United States could iteratively adapt recommendations in the NIST AI Risk Management Framework to incorporate current industry best practices. For the avoidance of doubt, these suggestions are not novel. Regulatory learning and adaptability are routine in the regulation of pharmaceuticals, aviation, and cybersecurity. AI governance should follow suit.

B. Diversification

> Principle 2: AI governance should adopt a portfolio approach comprised of diverse and uncorrelated regulatory strategies.

This principle, which advocates employing a range of heterogeneous strategies for governing AI, rests on two insights. The first is that the potential risks posed by AI are so great that policymakers cannot afford to put all their eggs in one basket. As in finance, they need to diversify their investments.[301]

COLUM. L. REV. 369 (2019); Rory Van Loo, *The Missing Regulatory State: Monitoring Businesses in an Age of Surveillance*, 72 VAND. L. REV. 1563 (2019).

[299] *See, e.g.*, Algorithmic Accountability Act §§ 3–5.

[300] For discussion of whistleblower protection, see Future of Life Inst., *Position Paper on the EU AI Act*, *supra* note 260, at 7; Clarke, Whittlestone, Maas, Belfield, Hernández-Orallo & Ó hÉigeartaigh, *supra* note 230, at 7. *See also* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 117–40 (2019); Hannah Bloch-Wehba, *The Promise and Perils of Tech Whistleblowing*, 118 NW. U. L. REV. (forthcoming). For discussion of financial incentives, including bug-, safety- and bias-bounties, see Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV at 16–18 (Apr. 20, 2020), https://arxiv.org/abs/2004.07213; Shahar Avin et al., *Filling Gaps in Trustworthy Development of AI*, 374 SCIENCE 1327, 1329 (2021); Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji & Joy Buolamwini, *Bug Bounties for Algorithmic Harms?*, ALGORITHMIC JUSTICE LEAGUE (Jan. 2022), https://www.ajl.org/bugs; OpenAI, *Announcing OpenAI's Bug Bounty Program*, OPENAI (Apr. 11, 2023), https://openai.com/blog/bug-bounty-program.

[301] The seminal economics paper on portfolio theory and diversification is Harry

Instead of relying on a single institutional framework, policymakers need to invest in a portfolio of risk mitigation measures. Ideally, these measures should not be correlated with one another, such that the failure of one will not necessarily lead to the failure of others.[302] The second insight is that the causes of algorithmic black swans are complex and multifaceted, arising in different stages of the AI value chain and influenced by the actions of multiple stakeholders.[303] Accordingly, policymakers should develop regulatory tools that target multiple sites of intervention.

What will diversified AI governance involve in practice? To begin with, policymakers should aim to address all parts of the AI value chain.[304] At present, far greater emphasis is placed on deployment in downstream applications, neglecting risks arising during research and development. For example, while several leading AI companies agreed to a set of best practices for safely deploying large language models,[305] no (publicly known) agreement has been reached on safety protocols for research and development.[306] Similarly, while there exists a database of safety incidents encountered in the deployment of automated systems,[307] there is no

---

Markowitz, *Portfolio Selection*, 7 J. FIN. 77 (1952), for which Markowitz was subsequently awarded the Nobel Memorial Prize in Economic Sciences in 1990.

[302] This risk mitigation strategy is sometimes described as "defense in depth." *See, e.g.*, *Defense in Depth*, NUCLEAR REGULATORY COMMISSION (Mar. 9, 2021), https://www.nrc.gov/reading-rm/basic-ref/glossary/defense-in-depth.html (describing the use of "multiple independent and redundant layers of defense to compensate for potential human and mechanical failures so that no single layer, no matter how robust, is exclusively relied upon."). *See also* Hendrycks, Mazeika & Woodside, *supra* note 14, at 30 (discussing a "Swiss cheese model" for multilayered organizational safety). These strategies are particularly useful in preventing cascading failures. *See generally* Sergey V. Buldyrev et al., *Catastrophic Cascade of Failures in Interdependent Networks*, 464 NATURE 1025 (2010); MELANIE MITCHELL, COMPLEXITY: A GUIDED TOUR 255–57 (2010).

[303] *See* Engler & Renda, *supra* note 222, at 2 (defining the AI value chain as the "process through which an individual AI system is developed and then put into use (or deployed)."). *See generally* Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, PROC. 2023 ACM CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1186 (2023); Jennifer Cobbe & Jatinder Singh, *Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges*, 42 COMPUT. L. & SEC. REV. 1, 3–7 (2021).

[304] *See* Gebru et al., *supra* note 209, at 5 (discussing the importance of regulating general purpose systems across their entire product cycle).

[305] *See* OpenAI, *Best Practices for Deploying Language Models*, OPENAI (June 2, 2022), https://openai.com/blog/best-practices-for-deploying-language-models/ (describing a joint initiative of Cohere, OpenAI, and AI21 Labs).

[306] *But see* Jonas Schuett et al., *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion*, ARXIV (May 11, 2023), https://arxiv.org/abs/2305.07153 (surveying fifty-one researchers in industry, academia, and civil society).

[307] See *AI Incidents Database*, PARTNERSHIP ON AI, https://partnershiponai.org/workstream/ai-incidents-database/; Sean McGregor, *Preventing Repeated Real World AI*

equivalent platform for reporting safety incidents encountered during AI research and development.[308] Diversified governance should begin by addressing these gaps and tackling the range of risks distributed across all parts of the technology's value chain.

Another aspect of diversified governance involves targeting different stakeholders. Beginning at the top of the AI value chain, policymakers could regulate organizations that provide resources or inputs for AI development.[309] For example, policymakers could require chip manufacturers and cloud computing providers to vet prospective customers or ensure they are compliant with appropriate safety standards. Next, policymakers could incentivize AI developers to invest a larger fraction of their resources in improving the safety of systems they build. This could be facilitated through a safety tax,[310] financial support (such as subsidies),[311] or other interventions. Finally, policymakers could mitigate risks in downstream deployment by requiring that AI developers install more robust safeguards against negligent and malicious uses of systems they build.[312] For example, rather than publicly release state-of-the-art systems in their entirety, developers could provide

---

Failures by Cataloging Incidents: The AI Incident Database, 35TH AAAI CONF. ON AI (2021).

[308] *But see* EU AI Act art. 60 (proposing a database that will contain specifications and other information relating to certain high-risk systems). *See also* Violet Turri & Rachel Dzombak, *Why We Need to Know More: Exploring the State of AI Incident Documentation Practices*, PROC. 2023 AAAI /ACM CONF. AI, ETHICS & SOC'Y 576 (2023).

[309] *See, e.g.*, Lennart Heim, *Compute Governance*, ALIGNMENT FORUM (Oct. 14, 2021), https://www.alignmentforum.org/s/bJi3hd8E8qjBeHz9Z/p/M3xpp7CZ2JaSafDJB. Recent proposals include Yonadav Shavit, *What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*, ARXIV (May 30, 2023), https://arxiv.org/abs/2303.11341; Jaime Sevilla, Anson Ho & Tamay Besiroglu, *Please Report your Compute*, 66 COMM. ACM 30 (2023); Hanna Dohmen, Jacob Feldgoise, Emily S. Weinstein & Timothy Fist, *Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policymakers, Part I*, GEORGETOWN CENTER FOR SECURITY & EMERGING TECHNOLOGY (May 15, 2023), https://cset.georgetown.edu/article/controlling-access-to-advanced-compute-via-the-cloud/; *Part II*, https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/.

[310] *See, e.g.*, Rui-Jie Yew & Dylan Hadfield-Menell *A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems*, PROC. 2022 AAAI /ACM CONF. AI, ETHICS & SOC'Y 823 (2022); Mckay Jensen, Nicholas Emery-Xu & Robert Trager, *Industrial Policy for Advanced AI: Compute Pricing and the Safety Tax*, ARXIV (Feb. 22, 2023), https://arxiv.org/abs/2302.11436.

[311] *See, e.g.*, Daniel E. Ho et al., *Building a National AI Research Resource: A Blueprint for the National Research Cloud*, STANFORD INSTIT. FOR HUMAN-CENTERED AI (Oct. 2021), https://hai.stanford.edu/white-paper-building-national-ai-research-resource.

[312] *See* Brundage et al., *Malicious Use*, *supra* note 232, at 50–57. Building effective safeguards, however, is notoriously difficult. *See, e.g.*, Alexander Wei, Nika Haghtalab & Jacob Steinhardt, *Jailbroken: How Does LLM Safety Training Fail?*, ARXIV (Jul. 5, 2023), https://arxiv.org/abs/2307.02483.

users with "structured access" whereby limits are placed on a system's use, modification, and reproduction.[313]

Importantly, diversified governance does not only require expanding the range of regulatory *targets*. It also requires expanding the range of *regulators*. Just as governance strategies focused on one regulatory target (e.g., software developers) can fail, governance strategies relying on one particular regulator can also fail. This problem is well known in financial regulation. Economist Jón Daníelsson makes the following plea:

> If we put a single regulator in charge of everything—the super regulator so common today—we end up with a government agency that prefers uniformity, one that shares the goals of the incumbent interests and loathes what is different. We need competition between regulators, so we get agencies that both regulate and defend their part of the industry, protecting heterogeneity along the way.[314]

The suggestion, in other words, is to *diversify the regulators*, and ideally introduce competition among them.[315] One compelling proposal for AI governance involves policymakers creating "regulatory markets" in which private sector organizations compete to achieve overarching governance goals.[316] Rather than engaging a single regulator to audit the safety of automated systems, policymakers could establish a framework in which multiple private auditing firms compete for the business of AI companies, leveraging the expertise and incentives of those firms to develop more rigorous and scalable safety audits.

---

[313] *See* Toby Shevlane, *Structured Access: An Emerging Paradigm for Safe AI Deployment*, ARXIV at 3 (Apr. 11, 2022), https://arxiv.org/abs/2201.05159. *See also id.* at 6 ("Structured access is rooted in a broader phenomenon, going beyond AI, where the owners of potentially harmful artefacts attempt to place limits on how users can interact with those artefacts. For example, certain biological laboratories have the capability to print DNA sequences and offer this as a service. The synthesized DNA can be used for beneficial research but could in theory be used for the creation of bioweapons.").

[314] JÓN DANÍELSSON, THE ILLUSION OF CONTROL: WHY FINANCIAL CRISES HAPPEN, AND WHAT WE CAN (AND CAN'T) DO ABOUT IT 252 (2022).

[315] *See* GILLIAN K. HADFIELD, RULES FOR A FLAT WORLD: WHY HUMANS INVENTED LAW AND HOW TO REINVENT IT FOR A COMPLEX GLOBAL ECONOMY 248, 265 (2017) (advocating the establishment of competitive markets of private regulators overseen by public authorities).

[316] *See* Clark & Hadfield, *supra* note 120, at 9 ("The key here is a shift by government to establishing the goals of regulation, rather than the methods of achieving those goals."). *See id.* at 8 (offering several concrete examples of applying this regime to AI systems).

C. Scalability

Principle 3: AI governance should be highly scalable.

Scalability describes the capacity of a system to function effectively with increasing workload.[317] While the term has its origins in the ability of computer systems to handle larger operational demands, scalability is often used to describe the ability of organizations to grow and adapt in the face of larger challenges and opportunities.[318] Scalability is also an important—though neglected—feature of regulation.[319] Scalable regulation describes regulation that continues to achieve its goals even as the organizations and systems with which it interacts increase in number and complexity.[320] Scalable regulation is particularly important in AI governance. To mitigate the risk of algorithmic black swans, policymakers will need to establish governance mechanisms that function effectively even as AI systems perform more complex tasks and are deployed in higher-stakes domains.[321]

At present, many of the AI governance proposals in the United States and the European Union are not highly scalable. For example, both the White House AI Bill of Rights and the EU AI Act mandate a large degree of human oversight, that is, engaging humans to oversee the operation of AI systems.[322] The resources required to meet this demand, especially if automated systems are deployed at large scale and operate at high speed, are prohibitive.[323]

---

[317] *See* André B. Bondi, *Characteristics of Scalability and Their Impact on Performance*, PROC. 2ND INT'L WORKSHOP ON SOFTWARE & PERFORMANCE 195 (2000).

[318] *See* Charles B. Weinstock & John B. Goodenough, *On System Scalability* (Technical Note, Carnegie Mellon University, Mar. 2006), https://resources.sei.cmu.edu/asset_files/TechnicalNote/2006_004_001_14681.pdf.

[319] *See* Cristie Ford, *Prospects for Scalability: Relationships and Uncertainty in Responsive Regulation*, 7 REG. & GOV. 14, 17–21 (2013).

[320] *Id.*

[321] To be clear, technical tools for improving the safety of AI systems must also be highly scalable. *See, e.g.*, Samuel R. Bowman et al., *Measuring Progress on Scalable Oversight for Large Language Models*, ARXIV at 1 (Nov. 11, 2022), https://arxiv.org/abs/2211.03540 ("To build and deploy powerful AI responsibly, we will need to develop robust techniques for scalable oversight: the ability to provide reliable supervision … to models in a way that will remain effective past the point that models start to achieve broadly human-level performance"); Amodei, Olah, Steinhardt, Christiano & Mané, *supra* note 14, at 3 (arguing that AI systems must operate safely even when it is impractical or impossible for a human to oversee their actions).

[322] *See* White House AI Bill of Rights, *supra* note 182, at 47 ("criminal justice system, employment, education, healthcare, and other sensitive domains … require extra protections. It is critically important that there is *extensive human oversight* in such settings." (emphasis added)). *See also* EU AI Act art. 14(1) ("High-risk AI systems shall be designed and developed in such a way … that they can be effectively overseen by natural persons").

[323] *See, e.g.*, Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans*

Moreover, mandating human oversight arguably undermines the societal benefits of many AI technologies. For example, requiring that AI systems that provide professional services remain subject to human oversight could increase the costs of these systems, such that the most significant beneficiaries—people who cannot afford traditional professional services—are priced out of the technology.[324]

The solution is not necessarily to allocate more resources to the problem, but to allocate the right resources. Scaling regulation to address the monumental challenges posed by AI requires us to shift away from relying on rigid written rules and cumbersome compliance mechanisms, and toward building regulatory technology that adapts to the evolving risks from AI.[325] Put simply, AI regulation "will require almost as much or more AI than the AI targets of regulation themselves."[326] Instead of mobilizing armies of bureaucrats to oversee high-risk automated decisions, policymakers should invest in technologists who innovate new, automated methods for auditing these decisions.[327] Some of these methods will surely fail, or even backfire. But others might succeed. Cautious experimentation is the only way to find out.

D. Experimentation

> Principle 4: AI governance should continually explore and evaluate new regulatory strategies.

It is wishful thinking to assume that the governance mechanisms currently proposed in the United States or the European Union are optimal. They represent only a small fraction of the possible strategies for regulating AI. The central problem is not the selection of certain governance strategies and priorities (and the exclusion of others), but the rigidity of the proposals themselves—that is, their inability to embrace new governance strategies or abandon existing ones.

---

in the Loop, 76 VAND. L. REV. 429, 455 (2023); Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, 45 COMPUT. L. & SEC. REV. 105681, 7–11 (2022).

[324] *See* Kolt, *supra* note 122, at 132–33. *Compare* Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 50–53 (2019).

[325] *See* HADFIELD, *supra* note 315, at 5. *See also* SULEYMAN, *supra* note 33, at ch. 14 ("Managing powerful tools itself requires powerful tools.").

[326] *Id.* at 247.

[327] Several approaches have been developed to accomplish "scalable oversight." *See, e.g.*, Samuel R. Bowman et al., *supra* note 321; Perez et al., *Discovering Language Model Behaviors*, *supra* note 15; Ethan Perez et al., *Red Teaming Language Models with Language Models*, PROC. 2022 CONF. EMPIRICAL METHODS IN NLP 3419 (2022).

For example, the White House AI Bill of Rights, the NIST AI Risk Management Framework, and the EU AI Act all prioritize algorithmic transparency.[328] However, it could turn out that transparency is technically unfeasible or practically unhelpful in high-stakes contexts.[329] At the same time, these proposals exclude potentially promising governance strategies, such as incentivizing technologists and members of the public to discover vulnerabilities in safety-critical AI systems.[330] To restate the problem, current regulatory proposals both entrench many untested strategies for governing AI and fail to establish mechanisms for exploring and incorporating new, possibly more effective, governance strategies.

Although there is no simple workaround, policymakers concerned about the prospect of algorithmic black swans would benefit from adopting a more experimental approach to AI governance. Just as experimentation is vital to scientific progress, venture capital investment, and product marketing, experimental techniques could provide policymakers with useful information about the best strategies for mitigating high-impact risks from AI.[331] Regulatory experimentation involves two steps: *exploration* and *evaluation*. In exploration, policymakers investigate and prototype a broad range of novel regulatory strategies.[332] In evaluation, policymakers subject these strategies (alongside existing regulatory strategies) to rigorous testing, assessing how they perform in practice.[333]

---

[328] *See* White House AI Bill of Rights, *supra* note 182, at 6; NIST AI Risk Management Framework, *supra* note 171, at 15–17; EU AI Act art. 13(1).

[329] For example, explanations regarding the operation of AI systems that generate computer code at high speed may do little to improve the safety of such systems. *See supra* Parts I.B, II.C.

[330] *See supra* note 300 (discussing the use of bug-, safety- and bias-bounties).

[331] On the use of experimental methods, including randomized trials, in lawmaking, see Michael Abramowicz, Ian Ayres & Yair Listokin, *Randomizing Law*, 159 U. PA. L. REV. 929, 934–38 (2011); Zachary J. Gubler, *Experimental Rules*, 55 B.C. L. REV. 129, 129–30 (2014); Wansley, *supra* note 260, at 432–36. *See also id.* at 404 (suggesting that "agencies are empowered to impose moratoria on risky emerging technologies while regulators organize experiments to learn about the risks they pose and the means to mitigate them.").

[332] *See* Ganguli et al., *supra* note 24, at 1757.

[333] *See, e.g.*, Cary Coglianese, *Empirical Analysis and Administrative Law*, 2002 U. ILL. L. REV. 1111, 1116–17 (2002) (surveying experimental and observational approaches to empirically studying the effect of regulatory interventions). *See also* Abramowicz, Ayres & Listokin, *supra* note 331, at 933 ("government should embrace randomized trials of statutes and regulations as a tool for testing the effectiveness of those laws. Just as random assignment of treatments is the most powerful method of testing for the causal impact of pharmaceuticals, random assignment of individuals, firms, or jurisdictions to different legal rules can help resolve uncertainty about the consequences of laws and regulations."). *Compare* Wansley, *supra* note 260, at 405 ("For many risks, from asbestos to climate change, the relevant science is settled, so there is little marginal value to publicly organized experiments. For other risks, especially catastrophic risks, randomized experiments might

Consider a concrete example. Policymakers, disappointed by the results of existing frameworks for reporting AI safety incidents,[334] design a bounty scheme whereby technologists and members of the public who discover vulnerabilities in high-risk automated systems receive financial rewards.[335] To evaluate the effectiveness of this scheme, policymakers compare the quality and quantity of safety information it uncovers to the information collected from existing reporting frameworks. Policymakers randomly assign bounty hunters to different groups, which receive different rewards, in order to discover which rewards most successfully elicit critical safety information. Finally, after finding that the bounty scheme offers some improvements over existing reporting frameworks, policymakers propose a stress test:[336] they deliberately insert vulnerabilities into several high-risk automated systems and observe whether any bounty hunters catch the bait.

This experimental process is foreign to most policymakers. Rather than establish once-and-for-all rules, the process tasks policymakers with iteratively designing and testing new governance strategies. It encourages the exploration and evaluation of novel regulatory tools. Throughout the process, policymakers receive high-quality feedback on their proposals and gain real-world insight into which governance strategies work in practice. Although costly and demanding, experimentation will be key to developing regulations that effectively anticipate and mitigate large-scale societal risks from AI.

E. Recalibrating Risk

> Principle 5: Cost-benefit analysis of AI governance interventions should place greater weight on worst-case outcomes.

Implementing the AI governance principles discussed above—by establishing a dynamic portfolio of diverse and scalable regulatory strategies—is a resource-intensive exercise. Algorithmic preparedness requires an unusual combination of foresight, technological expertise, and institutional flexibility. Successfully establishing the kind of governance structures envisaged in the principles above could also impose considerable social and economic costs. As with other regimes for governing emerging technologies, the regulation of AI could have a chilling effect on innovation, stifling progress in the field and denying society the tremendous gains offered

---

not be feasible or ethical. Some risks are latent for decades, so controlled experiments would take too long for any concurrent moratoria to be meaningfully temporary.").

[334] *See supra* note 223 (discussing Article 62 of the EU AI Act).

[335] *See supra* note 300 (describing the use of bug-, safety- and bias-bounties).

[336] *See generally* Rory Van Loo, *Stress Testing Governance*, 75 VAND. L. REV. 553 (2022) (examining the use of stress tests by government agencies).

by the technology.[337] How should policymakers navigate this tradeoff? How can they encourage the development of AI technologies that benefit society while curtailing the risk of potentially catastrophic outcomes?

The traditional answer, grounded in the process of cost-benefit analysis, is to weigh the prosocial benefits of proposed regulation (public health, welfare, safety, etc.) against the costs of such regulation (economic growth, innovation, competitiveness, etc.).[338] To the extent possible, policymakers strive to quantify these benefits and costs ahead of time. In the case of emerging technologies, however, this is exceedingly difficult. The most significant benefits and costs of AI regulation are likely to be "unknown unknowns,"[339] which cannot be quantified.[340] The question thus becomes: how should policymakers act in the face of uncertainty?

One approach, encapsulated in the precautionary principle, is to prioritize safety at all costs.[341] The principle supports placing stringent limitations on

---

[337] *See, e.g.*, Rebecca Janßen, Reinhold Kesler, Michael E. Kummer & Joel Waldfogel, *GDPR and the Lost Generation of Apps*, (NBER Working Paper No. 30028, May 2022), https://www.nber.org/papers/w30028 (finding that the GDPR led to a significant decrease in the number of apps available on Google Play and, in turn, detrimentally affected consumer choice and consumer welfare).

[338] *See* Exec. Order No. 13563 § 1(b), 76 Fed. Reg. 3821 (Jan. 18, 2011) ("Our regulatory system must protect public health, welfare, safety, and our environment while promoting economic growth, innovation, competitiveness, and job creation. … It must take into account benefits and costs, both quantitative and qualitative."); Holdren, Sunstein & Siddiqui, *supra* note 289, at 2 ("Federal regulation and oversight of emerging technologies should be based on an awareness of the potential benefits and the potential costs of such regulation and oversight"). *See generally* CASS R. SUNSTEIN, THE COST-BENEFIT REVOLUTION (2018). For influential critiques of cost-benefit analysis, see Frank Ackerman & Lisa Heinzerling, *Pricing the Priceless: Cost-Benefit Analysis of Environmental Protection*, 150 U. PA. L. REV. 1553 (2002); FRANK ACKERMAN & LISA HEINZERLING, PRICELESS: ON KNOWING THE PRICE OF EVERYTHING AND THE VALUE OF NOTHING (2004); Julie E. Cohen, *The Regulatory State in the Information Age*, 17 THEORETICAL INQ. L. 369, 392–96 (2016).

[339] *See* Rumsfeld, *supra* note 99; Kaminski, *supra* note 36; at 153; Crootof & Ard, *supra* note 131, at 380–81.

[340] *See* Exec. Order No. 13563 § 1(b), 76 Fed. Reg. 3821 (Jan. 18, 2011) ("recognizing that some benefits and costs are difficult to quantify"); OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, CIRCULAR A-4 (2003) ("In some cases, the level of scientific uncertainty may be so large that you can only present discrete alternative scenarios without assessing the relative likelihood of each scenario quantitatively.") These challenges are well known to scholars of regulation. *See* Farber, *supra* note 98, at 909; Cass R. Sunstein, *The Limits of Quantification*, 102 CAL. L. REV. 1369, 1373–85 (2014); Jonathan S. Masur & Eric Posner, *Unquantified Benefits and the Problem of Regulation under Uncertainty*, 102 CORNELL L. REV. 87 (2016). *See also* OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, DRAFT CIRCULAR A-4 at 11, 46, 81 (Apr. 6, 2023), https://www.whitehouse.gov/wp-content/uploads/2023/04/DraftCircularA-4.pdf (incorporating "catastrophic" risks into the framework of cost-benefit analysis).

[341] *See* Cass R. Sunstein, *Beyond the Precautionary Principle*, 151 U. PA. L. REV. 1003, 1003–4 (2003) (offering a pithy summary of the principle: "better safe than sorry."). For

potentially hazardous technologies.[342] The problem with this approach is that it does not place sufficient weight on the risks and opportunity costs of *not* using a technology.[343] Prosocial AI applications could, for example, deliver unprecedented economic and scientific gains.[344]

More importantly, the precautionary principle fails to recognize that both regulatory action and regulatory inaction can be costly.[345] For example, just as government intervention dramatically accelerated the development of COVID-19 vaccines, government intervention could spur much-needed innovation to improve the safety of high-risk AI systems.[346] At the same time, regulatory intervention can sometimes backfire, leading to undesirable unintended consequences. For example, public health researchers found that some pandemic lockdowns caused inadvertent harm by reducing access to healthcare services.[347] Similarly, the imposition of onerous restrictions on AI could hinder progress in developing tools that improve the technology's safety and social impact.

---

discussion of how the precautionary principle should shape AI regulation, see Kaminski, *supra* note 36, at 148–150.

[342] *See* Farber, *supra* note 98, at 914.

[343] *See* Crootof & Ard, *supra* note 131, at 385.

[344] *See*, *e.g.*, Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 SCIENCE, 187 (2023); Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond, *Generative AI at Work* (NBER Working Paper 31161, Apr. 2023), https://www.nber.org/papers/w31161; Ege Erdil, Tamay Besiroglu, *Explosive Growth from AI Automation: A Review of the Arguments*, ARXIV (Oct. 1, 2023), https://arxiv.org/abs/ 2309.11690; Daniil A. Boiko, Robert MacKnight & Gabe Gomes, *Emergent Autonomous Scientific Research Capabilities of Large Language Models*, ARXIV (Apr. 11, 2023), https://arxiv.org/abs/2304.05332; Gary Charness, Brian Jabarian & John A. List, *Generation Next: Experimentation with AI*, (NBER Working Paper No. 31679, Oct. 2023), https://www.nber.org/papers/w31679. *See also supra* note 205 (characterizing AI as a general purpose technology).

[345] *See* Cass R. Sunstein, *Maximin*, 37 YALE J. ON REGUL. 940, 951 (2020). *See also* Sunstein, *Beyond the Precautionary Principle*, *supra* note 341, at 1054; POSNER, *supra* note 93, at 140 (arguing that the precautionary principle "is not a satisfactory alternative to cost-benefit analysis, if only because of its sponginess—if it is an alternative at all."); Farber, *supra* note 98, at 916–19 (discussing several criticisms of the precautionary principle).

[346] *See* Moncef Slaoui & Matthew Hepburn, *Developing Safe and Effective Covid Vaccines: Operation Warp Speed's Strategy and Approach*, 383 NEW ENG. J. MED. 1701 (2020). For a comparable proposal in AI development, see Sethuraman Panchanathan & Arati Prabhakar, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resourc*e, NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE (Jan. 2023), https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

[347] *See* Constantin-Cristian Topriceanu et al., *Evaluating Access to Health and Care Services During Lockdown by the COVID-19 Survey in Five UK National Longitudinal Studies*, 11 BRIT. MED. J. OPEN (2021).

An alternative approach to confronting technological uncertainty is known as the *maximin* principle, which involves designing regulations that specifically address catastrophic risks.[348] According to Cass Sunstein, who offers the most detailed exploration of the principle, maximin instructs regulators to "choose the policy with the best worst-case outcome."[349] While Sunstein observes that the maximin principle is generally an inappropriate guide for crafting public policy,[350] he suggests that the principle is vital in some scenarios. Specifically, the principle is a useful guide in cases of Knightian uncertainty[351]—where potential risks cannot be assigned probabilities and conventional cost-benefit analysis cannot be undertaken—as is common in climate change, pandemics, and emerging technologies.[352] Given the inherent uncertainty around the risks (and benefits) from AI, policymakers will need to act with humility.[353] Neither regulatory action nor inaction is safe by default. Instead, policymakers should, in weighing the benefits and costs of governance strategies, focus on the overarching priority: protecting society from algorithmic black swans.

---

[348] *See* Farber, *supra* note 98, at 919 (explaining that maximin involves selecting the option that *maximizes* the *minimum* utilities across available options). For an accessible introduction, see CASS R. SUNSTEIN, AVERTING CATASTROPHE: DECISION THEORY FOR COVID-19, CLIMATE CHANGE, AND POTENTIAL DISASTERS OF ALL KINDS (2021).

[349] *See* Sunstein, *Maximin*, *supra* note 345, at 966.

[350] *Id.* at 943–44, 976.

[351] *See* FRANK H. KNIGHT, RISK, UNCERTAINTY, AND PROFIT 19–20 (1921) ("Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk … 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena …"). *See also* Farber, *supra* note 98, at 903 (discussing the ramifications of Knightian uncertainty: "Our society has sophisticated techniques for analyzing risks that can be modeled and quantified. But other threats—often the most serious ones—do not fit the paradigm.").

[352] *See* Sunstein, *Maximin*, *supra* note 345, at 950–51. Sunstein's analysis draws on insights from environmental economics. *See, e.g.*, Martin L. Weitzman, *Fat Tails and the Social Cost of Carbon*, 104 AM. ECON. REV. 544 (2014); *On Modeling and Interpreting the Economics of Catastrophic Climate Change*, 91 REV. ECON. & STAT. 1 (2009). The maximin principle could potentially justify counter-intuitive policy frameworks, such as an "inverse proportionality test" according to which more heavy-handed regulatory intervention is justified earlier (rather than later) in the unfolding of catastrophic events. *See* Ofer Malcai & Michal Shur-Ofry, *Using Complexity to Calibrate Legal Response to Covid-19*, 9 FRONTIERS IN PHYSICS 650943, 2–3 (2021) ("When the diffusion dynamics [of a pandemic] are nonlinear and the potential harm is likely to accumulate exponentially, strict measures to prevent it could be considered proportionate at an early stage, when the actual harm is least apparent and least certain. Counterintuitively, those very same measures might be less defensible at a later stage when the large harms of the pandemic have already materialized.").

[353] For a recent attempt to model this uncertainty, see Daron Acemoglu & Todd Lensman, *Regulating Transformative Technologies* (NBER Working Paper No. 31461, Jul. 2023), https://www.nber.org/papers/w31461.

CONCLUSION

The transformative impact of AI is only beginning to be felt. Eager to capture the benefits of the technology and combat the associated risks, policymakers in the United States and Europe are busy designing a host of new laws and policies that will shape the field in the coming decades. Many of these initiatives, however, overlook perhaps the most consequential challenge facing the governance of AI: mitigating large-scale societal harms. Without intervention, the risk of algorithmic black swans will compound as automated systems are deployed more widely and entrusted to perform increasingly important societal functions. Policymakers have a responsibility to anticipate and mitigate these risks. As in pandemic preparedness, no single intervention will suffice. Even a diverse portfolio of regulatory strategies may ultimately fail. This, however, does not undermine the case for preparing as best we can.

* * * * *