AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing[*]

Neel Guha,[†] Christie M. Lawrence,[†] Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, Daniel E. Ho[‡]

Stanford University

Abstract

Calls for regulating artificial intelligence (AI) are widespread, but there remains little consensus on both the specific harms that regulation can and should address and the appropriate regulatory actions to take. Computer scientists propose technical solutions that may be infeasible or illegal; lawyers propose regulation that may be technically impossible; and commentators propose policies that may backfire. AI regulation, in that sense, has its own alignment problem, where proposed interventions are often misaligned with societal values. In this Essay, we detail and assess the alignment and technical and institutional feasibility of four dominant proposals for AI regulation in the United States: disclosure, registration, licensing, and auditing. Our caution against the rush to heavily regulate AI without addressing regulatory alignment is underpinned by three arguments. First, AI regulatory proposals tend to suffer from both regulatory mismatch (i.e., vertical misalignment) and value conflict (i.e., horizontal misalignment). Clarity about a proposal's objectives, feasibility, and impact may highlight that the proposal is mismatched with the harm intended to address. In fact, the impulse for AI regulation may in some instances be better addressed by non-AI regulatory reform. And the more concrete the proposed regulation, the more it will expose tensions and tradeoffs between different regulatory objectives and values. Proposals that purportedly address all that ails AI (safety, trustworthiness, bias, accuracy, and privacy) ignore the reality that many goals cannot be jointly satisfied. Second, the dominant AI regulatory proposals face common technical and institutional feasibility challenges—*who* in government should coordinate and enforce regulation, *how* can the scope of regulatory interventions avoid ballooning, and *what* standards and metrics operationalize trustworthy AI values given the lack of, and unclear path to achieve, technical consensus? Third, the federal government can, to varying degrees, reduce AI regulatory misalignment by designing interventions to account for feasibility and alignment considerations. We thus close with concrete recommendations to minimize misalignment in AI regulation.

---

[*] This Essay is forthcoming in the *George Washington Law Review* Symposium on Legally Disruptive Emerging Technologies. We thank Stanford's Institute for Human-Centered Artificial Intelligence (HAI) for support, Rui-Jie Yew for research assistance, and Michael Abramowicz, David Freeman Engstrom, Alicia Solow-Niederman, Russell Wald, and attendees of the Symposium on Legally Disruptive Technologies for helpful comments.

[†] Equal authorship

[‡] Corresponding Author. Email: dho@law.stanford.edu.

Contents

I. Introduction

Announcing his company's scientific breakthrough, a tech CEO proclaimed, "This is clearly the first life form out of a computer and invented by humans."[1] This stunning research advance triggered a congressional hearing, intensive media coverage, and fears of a new form of "dual use" technology that could be used both to solve humanity's greatest challenges and create destructive bioweapons. With open online access to technology that could create synthetic genomes, could such technology enable "Do-It-Yourself" (DIY) biohacking, allowing any fringe individual to wreak havoc on the world?[2] Does such technology pose an existential threat to humanity by enabling the creation of novel pathogens outside of controlled laboratories? One article went so far as to posit that bioterrorists would be able to engineer a virus specifically targeted at the president's DNA.[3] While some called for the urgent need for regulation—for restricting access to scientific know-how to protect humanity—others warned against overreacting: "Do not overregulate something that needs care, integrity and responsibility."[4]

This debate was not about artificial intelligence (AI).[5] It was 2010 and the panic was about synthetic biology.[6] As the hype died down, doomsday scenarios failed to materialize, and the biohacking movement proved to be, at least for the moment, far more benign than either its proponents or opponents had believed. A Wilson Center study detailed not only how the vast majority of people involved in DIY Bio were still learning the basics of

[1] Maggie Fox, *U.S. Congress hears benefits of synthetic biology*, REUTERS (May 27, 2010), https://www.reuters.com/article/us-synthetic/u-s-congress-hears-benefits-of-synthetic-biology-idUKTRE64Q5YD20100527.
[2] Catherine Jefferson, Filippa Lentzos & Claire Marris, *Synthetic Biology and Biosecurity: Challenging the "Myths"*, 2 FRONTIERS IN PUBLIC HEALTH 115 (2014).
[3] Andrew Hessel, Marc Goodman & Steven Kotler, *Hacking the President's DNA*, ATLANTIC (Nov. 15, 2012), https://www.theatlantic.com/magazine/archive/2012/11/hacking-the-presidents-dna/309147/.
[4] Fox, *supra* note 1.
[5] One can easily find similar commentary about AI from policymakers and advocates today, however. *See, e.g.*, Press Release, Reps. Eshoo, Crenshaw Introduce Bill to Address AI Threats on Biosecurity (July 19, 2023), https://eshoo.house.gov/media/press-releases/reps-eshoo-crenshaw-introduce-bill-address-ai-threats-biosecurity.
[6] Jeanne Whalen, *In Attics and Closets, 'Biohackers' Discover Their Inner Frankenstein*, Wall Street Journal (May 12, 2009), https://www.wsj.com/articles/SB124207326903607931;
Carl Zimmer, *Amateurs Are New Fear in Creating Mutant Virus*, N.Y. TIMES (Mar. 5, 2012), https://www.nytimes.com/2012/03/06/health/amateur-biologists-are-new-fear-in-making-a-mutant-flu-virus.html.
Hanno Charisius, Richard Friebe & Sascha Karberg, *Becoming Biohackers: The Long Arm of the Law*, BBC (Jan. 23, 2013), https://www.bbc.com/future/article/20130124-biohacking-fear-and-the-fbi

biotechnology, but also that a culture of openness and transparency made infiltration by bad actors highly unlikely.[7]

With concerns brewing around existential risk,[8] bioweapons,[9] and terrorism,[10] the tenor of the AI debate bears an uncanny resemblance to the synthetic biology panic. One unpublished study by MIT researchers made the media rounds[11] for asserting that large language models (LLMs) could enable individuals with little knowledge (undergraduates spending an hour with models) to create the next pandemic.[12] If true, such reports are certainly cause for concern. Given the proclivity to regulate "dread risk,"[13] these reports have contributed wide-ranging proposals for regulation to (a) *stop* the development of LLMs;[14] (b) *ban or restrict* open[15] LLMs above a certain capacity;[16]

---

[7] DANIEL GRUSHKIN ET AL., SEVEN MYTHS & REALITIES ABOUT DO-IT-YOURSELF BIOLOGY (2013), https://www.wilsoncenter.org/sites/default/files/media/documents/publication/7_myths_final.pdf.
[8] *See, e.g.,* Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut it All Down*, TIME MAGAZINE (Mar. 29, 2023, 6:01 PM), https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.
[9] Jonas Sandbrink, *ChatGPT Could Make Bioterrorism Horrifyingly Easy*, VOX (Aug. 7, 2023, 7:00 AM), https://www.vox.com/future-perfect/23820331/chatgpt-bioterrorism-bioweapons-artificial-inteligence-openai-terrorism.
[10] António Guterres, Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence (July 18, 2023), *available at* https://press.un.org/en/2023/sgsm21880.doc.htm.
[11] *See, e.g.,* Robert F. Service, *Could Chatbots Help Devise the Next Pandemic Virus?*, SCIENCE MAGAZINE (June 14, 2023, 6:05 PM), https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus; Kelsey Piper, *How AI Could Spark the Next Pandemic*, VOX (June 21, 2023, 2:40 PM), https://www.vox.com/future-perfect/2023/6/21/23768810/artificial-intelligence-pandemic-biotechnology-synthetic-biology-biorisk-dna-synthesis; Sarah Newey and Paul Nuki, *Could AI chatbots be used to develop a bioweapon? You'd be surprised*, TELEGRAPH (July 6, 2023, 9:12 AM), https://www.telegraph.co.uk/global-health/science-and-disease/chatgpt-google-bard-ai-bioweapon-pandemic/.
[12] Emily H. Soice et al., *Can Large Language Models Democratize Access to Dual-Use Biotechnology?* 1, ARXIV (2023), https://arxiv.org/abs/2306.03809 ("[T]he 'Safeguarding the Future' course at MIT tasked non-scientist students with investigating whether LLM chatbots could be prompted to assist non-experts in causing a pandemic. In one hour, the chatbots suggested four potential pandemic pathogens, explained how they can be generated from synthetic DNA using reverse genetics, supplied the names of DNA synthesis companies unlikely to screen orders, identified detailed protocols and how to troubleshoot them, and recommended that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organization. Collectively, these results suggest that LLMs will make pandemic-class agents widely accessible as soon as they are credibly identified, even to people with little or no laboratory training.").
[13] Paul Slovic, *Perception of Risk*, 236 SCIENCE 280, 283 (1987).
[14] Yudkowsky, *supra* note 8.
[15] We note an ongoing debate regarding whether certain models can be described as "open source" or merely "open." *See* David Gray Widder et al., *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI*, SSRN (Aug. 18, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.
[16] *See, e.g.,* Press Release, Josh Hawley, Hawley and Blumenthal Demand Answers from Meta, Warn of Misuse After 'Leak' of Meta's AI Model (June 6, 2023), https://www.hawley.senate.gov/hawley-and-blumenthal-demand-answers-meta-warn-misuse-after-leak-metas-ai-mode.

(c) mandate *registration* of LLMs with penalties for non-registered use;[17] an (d) require a *license* to operate LLMs.[18]

Will such efforts reduce the risk of bioweapon development? Despite the headline-grabbing claim, the precise marginal risk of bioweapons manufacturing from LLMs is still unclear, given that many models may not do much more than regurgitate materials readily available on the internet or in library volumes.[19] As the Appendix illustrates, browsing *Wikipedia* yields pointers substantially similar to the MIT paper for how one might create the next pandemic.[20] And smaller non-LLMs can, just as well, predict novel toxic chemical compounds.[21] Without a detailed assessment of the capabilities of LLMs relative to other technologies, focusing on LLMs for bioweapons nonproliferation risks a mismatch between the object of the regulatory regime (limiting the development and use of LLMs) and the harm intended to be mitigated (catastrophic risk).[22]

The bioweapons example highlights two central questions for AI regulation: (1) whether regulatory compliance will in fact have a reasonable likelihood of materially mitigating the targeted harm at a feasible cost, and (2) whether compliance is even feasible. In this Essay, we argue that regulatory compliance must be front and center when conceiving of

[17] *See, e.g.,* Press Release, European Parliament, EU AI Act: first regulation on artificial intelligence (June 14, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[18] *See, e.g.,* Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html; Jeremy Kahn, *Microsoft: Advanced A.I. models need government regulation, with rules similar to anti-fraud and terrorism safeguards at banks*, FORTUNE MAGAZINE (May 12, 2023, 11:48 AM), https://fortune.com/2023/05/25/microsoft-president-says-the-u-s-must-create-an-a-i-regulatory-agency-with-rules-for-companies-using-advanced-a-i-models-similar-to-anti-fraud-safeguards-at-banks/.

[19] *See, e.g.,* NAT'L ACADS. OF SCIS., ENG'G & MED., BIODEFENSE IN THE AGE OF SYNTHETIC BIOLOGY (2018). Writes one law review piece, "[a]nyone seeking to design or manipulate pathogens can obtain the necessary tools to do so from commercial manufacturers in a number of ways." Braden Leach, *Necessary Measures: Synthetic Biology & the Biological Weapons Convention*, 25 STAN. TECH. L. REV. 141, 141 (2021).

[20] In any case, the fact that an LLM yields seemingly convincing answers does not mean that these answers are grounded in reality, given the extensively documented tendency of LLMs to "hallucinate" false information. *See, e.g.*, Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM COMPUTING SURVEYS 248:1, 248:2 (2023).

[21] *See* Fabio Urbina et al., *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTELLIGENCE 189 (2022) ("In less than 6 hours after starting on our in-house server, our model generated 40,000 molecules that scored within our desired threshold [of toxicity to humans]. In the process, the AI designed not only VX, but also many other known chemical warfare agents…").

[22] Put differently, which of the following may be more likely by 2024: more (a) open-source models, (b) laboratories capable of manufacturing pathogens, or (c) suppliers of required raw materials? If the answer is (a), the focus on (b) and (c) may provide more effective mechanisms of control. Others have written about the regulatory gaps in the control of bioweapons. *See* Leach, *supra* note 19.

regulatory interventions.[23] We argue that the optimal design of AI regulation is fundamentally different when technical and institutional constraints, both critical to compliance, are considered. Failure to do so will risk, at best, regulation as window dressing—and at worst, counterproductive or perverse downstream consequences. While more of our analysis focuses on the United States, this framework and its implications for AI regulation have applicability globally. We also cabin discussions of political feasibility (i.e., the ability of Congress to enact necessary legislation or regulators to navigate political constraints) to focus this Essay on regulatory design and enforcement. This is an important caveat, as regulatory design decisions in the real world may reflect policymakers' efforts to implement a potentially useful yet imperfect regulatory scheme while navigating a variety of political constraints.

We analyze compliance through the lens of *technical feasibility*–the availability of consensus technical and engineering solutions necessary to implement a regulatory proposal. A regulatory goal may be, at present, unachievable because it requires technology which does not currently exist. For instance, many proposals focus on disclosure of generative AI outputs through watermarking (i.e., identifying AI-generated output by inserting digital signatures or other specialized mechanisms into AI-produced output), but the ability to reliably watermark AI outputs is heavily disputed, particularly for text.[24] Regulatory interventions may also be frustrated by the fact that certain goals—like fairness[25]—lend themselves to diverse technical interpretations, which can often be in tension with each other.[26] Regulatory interventions which fail to acknowledge or account for such variation can induce confusion and inconsistency. Finally, even where the

[23] We borrow here from CYNTHIA GILES, NEXT GENERATION COMPLIANCE: ENVIRONMENTAL REGULATION FOR THE MODERN ERA (2022) (emphasizing the importance of designing environmental regulations "with compliance built in").

[24] Peter Henderson, *Should the United States or the European Union Follow China's Lead and Require Watermarks for Generative AI?*, GEO. J. FOR INT'L AFFS. (May 24, 2023), https://gjia.georgetown.edu/2023/05/24/should-the-united-states-or-the-european-union-follow-chinas-lead-and-require-watermarks-for-generative-ai/; *see also* Keith Collins, *How ChatGPT Could Embed a 'Watermark' in the Text It Generates*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/interactive/2023/02/17/business/ai-text-detection.html; Melissa Heikkilä, *A watermark for chatbots can expose text written by an AI*, MIT TECHNOLOGY REVIEW (Jan. 27, 2023), https://www.technologyreview.com/2023/01/27/1067338/a-watermark-for-chatbots-can-spot-text-written-by-an-ai/. AI detection tools like GPTZero and AI Classifier have also been shown to be inaccurate and even biased against non-native English speakers. *See* Benji Edwards, *OpenAI Confirms that AI Writing Detectors Don't Work*, ARS TECHNICA (Sept. 8, 2023, 11:42 AM), https://arstechnica.com/information-technology/2023/09/openai-admits-that-ai-writing-detectors-dont-work/; Weixin Liang et al., *GPT Detectors are Biased Against Non-native English Writers*, 4 PATTERNS 1 (2023).

[25] See *infra* note 68 and discussion in SectionII.B.

[26] Much of this debate has centered on how values like bias, privacy, and toxicity lend themselves to multiple computational interpretations, with little consensus as to which version should be adopted. For results showing the impossibility of satisfying certain definitions of fairness simultaneously, see Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, PROCS. OF THE 8TH CONF. ON INNOVATIONS IN THEORETICAL COMPUT. SCI. 43 (2017).

technology exists to implement an intervention, existing technical methods may nonetheless force value tradeoffs. Calls for more privacy-preserving AI, for instance, can conflict with calls for reducing algorithmic discrimination.[27] Proposals requiring all AI systems to produce explanations alongside predictions invoke all three types of technical infeasibility: existing methods (1) struggle to produce explanations for modern state-of-the-art AI systems, (2) fail to address technical disagreements about methods, and (3) may reduce model accuracy.[28]

In addition, a compliance-oriented perspective necessarily must grapple with each proposal's *institutional feasibility*, by which we mean the executive branch's institutional capacity to develop and effectively implement. For instance, calls for AI audits quickly run into major institutional challenges.[29] There is currently no agency well-positioned or resourced to conduct AI audits. Relying on audits conducted by parties external to the government requires trusting the independence of the auditors and accuracy of their audit—both notoriously difficult.[30]

Our Essay proceeds as follows. Section I discusses the wide range of harms AI regulation is thought to address. Sections II, III, IV, and V discuss four common proposals for AI regulation: the disclosure of AI system properties, registration of AI models or actors,[31] licensing of AI models or actors, and auditing of AI systems. For each proposal, we analyze the technical and institutional feasibility of the proposals, articulate how a focus on compliance should inform their design, and discuss how each proposal illustrates AI regulation's alignment problem. We focus on broader legislative proposals for AI regulation, noting that recent executive actions (e.g., the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence) include related interventions.

Our analysis cautions against immediately adopting heavy regulation of AI writ large without serious consideration of regulatory alignment and yields five themes discussed in greater length in Section VI. First, the four dominant I regulatory proposals face similar

---

[27] Alice Xiang, *Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision*, 36 HARV. J. L. & TECH. 1 (2022).
[28] See *infra* notes 128–132 and accompanying text.
[29] "Governments could legally require developers [to] provide model access pre-deployment to *government auditors*." Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, et al., Open Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives, Centre for the Governance of AI 22 (Sept. 29, 2023), https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models (emphasis added).
[30] *See infra* SectionVI.B (discussion of auditing's institutional feasibility).
[31] Actors may encompass the entities or individuals responsible for creating and training AI models, or those that use AI systems for certain applications.

technical and institutional feasibility challenges. Second, proposals may be mismatched with the risks intended to mitigate. Some risks associated with AI models may expose gaps in existing regulatory regimes that are better addressed by non-AI-focused regulation. Third as regulatory interventions become more concrete, they will increasingly reveal conflicts between heterogeneous goals of AI regulation that cannot be jointly satisfied.[32] Fourth, some regulation proposals could—even if potentially useful in advancing legitimate public objectives—function to advantage powerful incumbents in AI and reduce competition, thus stymieing innovation and concentrating AI's benefits.[33] Last, while textbook regulation is often predicated on categories of interventions (e.g., licensing vs. disclosure),[34] our analysis illustrates the malleability of conventional categories. However, the federal government can reduce the AI regulatory misalignment. We close by encouraging policymakers to focus on regulatory interventions that address current information asymmetries about emergent risks posed by AI (e.g., with adverse event reporting), explore institutional mechanisms for oversight of third-party audits, avoid the impulse to create a new super-agency for AI, and refrain from grappling with value tradeoffs by assuming non-governmental entities can easily operationalize technically feasible and value-neutral AI principles.

While scholars and citizens alike have bemoaned the inefficiency that seems to plague bureaucratic institutions, well-designed policies can mitigate organizational challenges. "American public bureaucracy is not designed to be effective,"[35] and unless policymakers take seriously the technical and institutional feasibility of their proposals, neither will AI regulation.

II. AI Regulation's (Mis)Alignment Problem

Effective and clear regulation requires clarity about the nature of the harm (or market failure) a regulation is seeking to address. In this section, we first articulate the kaleidoscopic nature of posited AI harms and then discuss what we call the "regulatory alignment problem."[36]

---

[32] *Cf.* Mark A. Lemley, *The Contradictions of Platform Regulation*, 1 J. FREE SPEECH L. 303 (2021).
[33] *See supra* note 48.
[34] *Id*.
[35] Terry Moe, *The Politics of Bureaucratic Structure*, *in* CAN THE GOVERNMENT GOVERN? 267, 267 (J. E. Chubb & P. E. Peterson eds., 1989).
[36] The "regulatory alignment problem" plays upon the broader AI alignment problem, which is "the idea that AI systems' goals may not align with those of humans, a problem that would be heightened if superintelligent AI systems are developed." Eliza Strickland, OpenAI's Moonshot: Solving the AI Alignment Problem, IEEE (Aug. 31, 2023), https://spectrum.ieee.org/the-alignment-problem-openai. AI misalignment is often a concern raised by those who are concerned that AI poses existential risks to humanity. *See, e.g.*, Jan Leike, What is the alignment problem? (Mar. 29, 2022), https://aligned.substack.com/p/what-is-alignment.

A. Calls to Regulate AI Emanate from Many Conceptions of Harm and Market Failure

Calls for regulation are predicated on a dizzying array of potential harms.[37] AI systems may exhibit poor performance[38] or declining performance over time or when applied in new contexts;[39] create or worsen disparities between demographic groups (i.e., bias);[40] contribute to surveillance[41] and the violation of information privacy.[42] AI systems can cause labor displacement[43] and the degradation of job quality.[44] AI systems have large

[37] Small excerpts in this section are derived from NAIAC EXEC. ACTION & REGULATION WORKING GRP., RATIONALES, MECHANISMS, AND CHALLENGES TO REGULATING AI: A CONCISE GUIDE AND EXPLANATION (2023), https://www.ai.gov/wp-content/uploads/2023/07/Rationales-Mechanisms-Challenges-Regulating-AI-NAIAC-Non-Decisional.pdf, which one of the authors drafted.

[38] Poor performance by AI systems has many causes. *See, e.g.*, *The Effects of Data Quality on Machine Learning Performance* (arXiv, Nov. 9, 2022) (low-quality or insufficient training data); Inioluwa Deborah Raji et al., *The Fallacy of AI Functionality*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2022) (poor suitability for a given domain). Failure modes vary widely and include hallucination of false information, Ji, *supra* note 20, generation of insecure computer code, Neil Perry et al., *Do Users Write More Insecure Code with AI Assistants?*, ARXIV (Dec. 16, 2022), https://arxiv.org/abs/2211.03622, and erratic behavior in interactions with users, Kevin Roose, *A Conversation With Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 16, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html ("The version [of Bing Chat] I encountered seemed… like a moody, manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine.").

[39] *See* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models* 109–113, ARXIV (July 12, 2022), https://arxiv.org/abs/2108.07258 ("High-stakes applications... require models that generalize well to circumstances not seen in the training data, e.g., test examples from different countries, under different driving conditions, or from different hospitals. Prior work has shown that these types of distribution shifts can cause large drops in performance even in state-of-the-art models.").

[40] Algorithmic bias has been documented across many different domains in both the public and private sectors. *See, e.g.,* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. (racial and gender bias in facial analysis system); David Arnold et al., *Measuring Racial Discrimination in Algorithms*, 111 AEA PAPERS & PROC. 49 (racial bias in bail algorithm); Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, REUTERS (Oct. 10, 2018, 4:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (gender bias in resume review system).

[41] Steven Feldstein, *The Global Expansion of AI Surveillance*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Sep. 17, 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

[42] *See* Cameron F. Kerry, *Protecting Privacy in an AI-Driven World*, BROOKINGS INSTITUTE (Feb. 10, 2020), https://www.brookings.edu/articles/protecting-privacy-in-an-ai-driven-world/ [https://perma.cc/WL78-PDAE].

[43] *See* Tyna Eloundou et al., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, ARXIV (Aug. 22, 2023), https://arxiv.org/abs/2303.10130; *see generally* Daron Acemoglu & Pascual Restrepo, *Automation and New Tasks: How Technology Displaces and Reinstates Labor*, 33 J. ECON. PERSPECTIVES, Spring 2019, 3 (2019).

[44] KAREN LEVY, DATA DRIVEN: TRUCKERS, TECHNOLOGY, AND THE NEW WORKPLACE SURVEILLANCE (2022).

environmental footprints to train and operate.[45] AI may undermine cybersecurity[46] or be vulnerable to exploitation;[47] contribute to the industrial concentration of wealth and influence;[48] shift geopolitical power to foreign adversaries;[49] contribute to democratic erosion;[50] and cause catastrophic or existential risk to humanity.[51] Table 1 provides illustrative examples of how each of these risks can manifest in practice but is far from exhaustive.

[45] *See, e.g.,* Payal Dhar, *The Carbon Impact of Artificial Intelligence*, 2 NATURE MACH. INTELLIGENCE 423 (2020); *but see, e.g.*, Bill Tomlinson et al., *The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans*, ARXIV (Mar. 8, 2023) ("We find that an AI writing a page of text emits 130 to 1500 times less CO2 than a human doing so. Similarly, an AI creating an image emits 310 to 2900 times less.").

[46] S*ee, e.g.*, IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY: PROCEEDINGS OF A WORKSHOP (Anne Johnson & Emily Grumbling eds., 2019), https://doi.org/10.17226/25488 [hereinafter Johnson & Grumbling]; Perry et al., *supra* note 38.

[47] AI systems may be vulnerable to several forms of exploitation once deployed, including circumvention of safety restrictions. *See, e.g.*, Rohan Goswami, *ChatGPT's 'Jailbreak' Tries to Make the A.I. Break Its Own Rules, or Die*, CNBC (Nov. 18, 2019), https://www.cnbc.com/2023/02/06/chatgpt-jailbreak-forces-it-to-break-its-own-rules.html; Seyed-Mohsen Moosavi-Dezfooli et al., *DeepFool: A Simple and A Method to Fool Deep Neural Networks*, 2016 IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION 2574 (2016). Another form of exploitation seeks to modify the behavior of an AI system by "poisoning" the data on which it is trained. *See* Fahri Anıl Yerlikaya & Şerif Bahtiyar, *Data Poisoning Attacks Against Machine Learning Algorithms*, 208 EXPERT SYSTEMS WITH APPLICATIONS 118101 (2022).

[48] *See* Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power*, N.Y. TIMES (Sep. 26, 2019), https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html. ("The danger [of increasing compute needs], [computer scientists] say, is that pioneering artificial intelligence research will be a field of haves and have-nots. And the haves will be mainly a few big tech companies like Google, Microsoft, Amazon and Facebook, which each spend billions a year building out their data centers."); Jai Vipra & Anton Korinek, *Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT*, BROOKINGS INST. (Sep. 7, 2023), https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/.

[49] *See, e.g.,* Nat'l Sec. Comm'n on A.I., *2021 Final Report* (2021), https://www.nscai.gov/2021-final-report/.

[50] Use of AI systems to create and spread misinformation (such as "deep fake" images and videos) may be used to undermine particular candidates for election or trust in democratic institutions in general. *See* Maria Pawelec, *Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions*, 1 DIGITAL SOCIETY 19 (2022); Jackson Cote, *Deepfakes and Fake News Pose a Growing Threat to Democracy, Experts Warn*, NORTHEASTERN GLOBAL NEWS (Apr. 1, 2022), https://news.northeastern.edu/2022/04/01/deepfakes-fake-news-threat-democracy/ [https://perma.cc/THQ8-Z53C].

[51] *See* NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014) (discussing several hypothetical scenarios in which a superintelligent AI system could pose an existential risk to humanity).

Harm	Example
Poor performance and inaccuracy	Software engineers may rely on a code-generating AI that produces bug-ridden computer code.
Bias	An algorithm that recommends whether a defendant should be granted bail may treat Black defendants disproportionately harshly.
Surveillance and privacy invasion	AI-powered surveillance may be used to monitor and punish dissidents at a scale not previously feasible. AI may also be used to generate explicit content depicting individuals without their consent.
Labor displacement and job degradation	AI may automate substantial portions of many jobs, with an outsized impact on high-paying knowledge work.
Environmental costs	Training of a large language model can create as much as 300,000 kg of carbon dioxide emissions, the equivalent of 125 round-trip flights from New York to Beijing.
Security	AI systems may discover exploits in computer systems or engage in social-engineering attacks against people with access to critical systems.
Concentration of industrial power and anti-competitive behavior	A small number of large corporations may control the best performing AI systems and capture AI's economic benefits at the expense of others.
Geopolitical power shift	Adversaries may advance AI capabilities faster than the U.S. and gain military or economic superiority.
Democratic erosion	AI may be used to create disinformation for dissemination online that undermines a candidate for political office.
Catastrophic risk	An advanced AI system may be used to design a bioweapon that could cause a global pandemic.

Table 1: The wide range of contemplated AI harms that animate different regulatory proposals.[52]

B. Proposals to Regulate AI Suffer from the Regulatory Alignment Problem

[52] For poor performance and inaccuracy, see Perry et al., *supra* note 38. For bias, see Arnold et al., *supra* note 40. For AI-powered surveillance monitoring dissidents, see Feldstein, *supra* note 41; Kerry, *supra* note 42. For generating content without consent, see Nina Jankowicz, *I Shouldn't Have to Accept Being in Deepfake Porn*, ATLANTIC (June 25, 2023), https://www.theatlantic.com/ideas/archive/2023/06/deepfake-porn-ai-misinformation/674475/. For labor impacts, see Eloundou, *supra* note 43. For environmental costs, see Dhar, *supra note* 45. For security see Johnson & Grumbling, *supra* note 46. For concentration of economic power, see Lohr, *supra* note 48; Vipra & Korinek, *supra* note 48. For geopolitical power shifts, see Nat'l Sec. Comm'n on A.I., *2021 Final Report*, *supra* note 49. For democratic erosion, see Pawelec, *supra* note 50. For catastrophic risk, see Bostrom, *supra* note 51.

Many calls for regulation have been inspired by concerns about AI's alignment problem, which in its simplest form is the concern that an AI system may not advance human goals, values, and ethical principles.[53] How can we ensure that an AI system is sufficiently *aligned* with human values? Such misalignment can occur between intended human *values* and the model *objective* or between the model *objective* and model *behavior*.[54] In a commonly referenced parable, a CEO is upset that a shortage of paperclips undermines productivity (the value) and commands the design of an AI system to maximize the number of paperclips (the objective).[55] The paperclip maximizer is so powerful (Artificial General Intelligence, or "AGI") that it kills humans, including the CEO to obtain more material for paperclip production (behavior). The objective of more paperclips is not perfectly aligned with the underlying human value of productivity, and the perverse behavior of the paperclip maximizing AI system is certainly each misaligned with productivity. While the alignment problem—and portrayal of AGI's existential risk to humanity—is used to illustrate the need for regulation, AI regulation suffers from its own alignment problem. What we term the "regulatory alignment problem" has two components: (a) *regulatory mismatch*—the fact that values may be misaligned with regulatory objectives and with behavior resulting from the regulatory system; and (b) *value conflict*—unrecognized tension between values that may require tradeoffs (e.g., the tradeoff between informational privacy and bias assessment and mitigation).

Table 2 illustrates the AI regulatory alignment problem by example. The left column depicts the conventional AI alignment problem of the paperclip maximizer. The right three columns depict the regulatory alignment problem with three distinct notions of AI harms: privacy violations, bias, or catastrophic risk.

[53] *See supra*, note 36; Blair Levin et al., *Who is Going to Regulate AI?*, HARV. BUS. REV. (May 19, 2023), https://hbr.org/2023/05/who-is-going-to-regulate-ai. We avoid a detailed discussion of the AI alignment problem for simplicity.
[54] This former is commonly referred to as the "outer alignment" problem and the latter as the "inner alignment" problem. Evan Hubinger et al., *Risks from Learned Optimization in Advanced Machine Learning Systems*, ARXIV (Jun. 5, 2019), https://arxiv.org/abs/1906.01820.
[55] Kathleen Miles, *Artificial Intelligence May Doom The Human Race Within A Century, Oxford Professor Says*, HUFFPOST (Feb. 4, 2015), https://www.huffpost.com/entry/artificial-intelligence-oxford_n_5689858.

	AI Alignment	Regulatory Alignment		
Observed Risk or Market Failure	Insufficient paperclips	Release of personally identifiable information (PII)	Disparities in hiring	Release of bioweapon construction information
Human Value	Productivity	Privacy	Fairness	Safety
Model / Regulatory Objective	Maximize paperclips	Differential privacy	80% Rule	Restriction of large language models
Unintended consequence (behavior)	Kill humans for paperclip material	Configure algorithms, given imprecise guidance, to aggressively mine data without protecting PII	Discard feature most predictive of job performance, decreasing accuracy and the "fairness" of the model	Use smaller, proprietary models to access sensitive information about bioweapons removing visibility into proliferation risk and preventing the identification of gaps in regulatory regimes (e.g., insufficient lab safety)

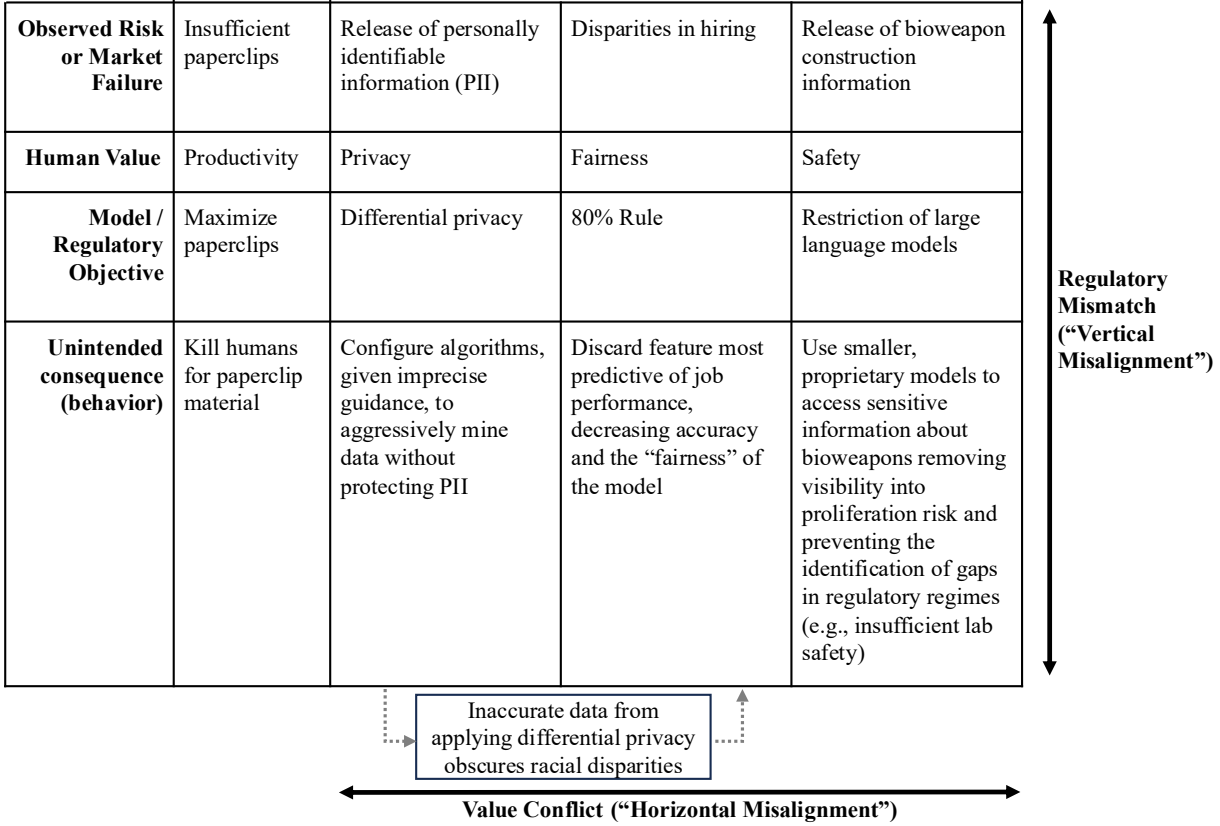

Table 2: The regulatory alignment problem. The left column depicts the conventional AI alignment problem with misalignment between the human value and (a) the model objective and/or (b) the model behavior. The right two columns illustrate the AI regulatory alignment problem—both vertical and horizontal misalignment.[56]

[56] On differential privacy and the unintended consequences, see Cynthia Dwork et al., *The Algorithmic Foundations of Differential Privacy*, 9 FOUNDATIONS AND TRENDS IN THEORETICAL COMP. SCI. 211 (2014); Andy Greenberg, *How One of Apple's Key Privacy Safeguards Falls Short*, WIRED (Sep. 15, 2017), https://www.wired.com/story/apple-differential-privacy-shortcomings/.Alexis R. Santos-Lozada et al., *How Differential Privacy Will Affect Our Understanding of Health Disparities in the United States*, 117 PROC. NAT'L ACAD. SCIS. 13405 (2020). On the 80% Rule and how features predictive of performance may be discarded, see Elizabeth Anne Watkins et al., *The Four-Fifths Rule Is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness* 1, ARXIV (Feb. 19, 2022), https://arxiv.org/pdf/2202.09519.pdf; Michael Feldman et al., *Certifying and Removing Disparate Impact*, PROC. 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 259 (2015). The MIT paper expressed concern that and LLM embedded into a chatbot suggested four potential pathogens, but concerns that AI models like AlphaFold could be dual-use technologies weaponized to identify harmful pathogens and proteins were already present. Emily H. Soice et al., *supra* note 14; Chris Miller, *There's a New US National Security Obsession — Biotech*, FIN. TIMES (Mar. 6, 2023), https://www.ft.com/content/cb9cd845-e9b0-4243-97f3-c315dac11fb4; Ying-Chiang J. Lee, Alexis Cowan, & Amari Tankard, *Peptide Toxins as Biothreats and the Potential for AI Systems to Enhance Biosecurity*, FRONT BIOENG. BIOTECH. (2022), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8959115/; Sterlin Sawaya, Taner Kuru, Thomas A. Campbell, *The Potential For Dual-Use of Protein-Folding Prediction*, UNICRI (2020), https://f3magazine.unicri.it/?p=2307.

Regulatory Proposals may be Mismatched with the Intended Harm Reduction ("Vertical Misalignment"). Regulatory interventions are most effective when tailored to address the underlying problem,[57] but proponents of regulation can be wildly imprecise about which harm(s) their proposed regulatory mechanism purports to address. And the relative importance of harms—and magnitude of harms relative to those imposed by non-AI baseline systems—can be fiercely contested. Thus, the required severity of a regulatory mechanism may be contentious. In regulatory theory, this problem has long been dubbed one of "regulatory mismatch," and we can conceive of it as tension between cells within a column (or also "vertical misalignment"). In short: how well does an intervention actually address the harm regulators seek to remediate?

Regulatory Mismatch Between the Observed Risk and the Desired Values and Regulatory Objectives of the Proposal. To state the obvious, achieving AI-related regulatory and policy goals requires tailoring the proposal to address the harm. If the concern is one of environmental costs, for instance, a typical intervention might be to tax energy-intensive computing (to incentivize parties to internalize the pollution cost[58]). Similarly, if the concern is about existential risk, an intervention might focus on restricting access generally to compute [59]). However, if regulators were concerned about the barriers to entry for AI development and national competitiveness more broadly, then a natural intervention might be to subsidize compute[60] to spur more market entrants.[61] In its simplest form, mismatch occurs if a proposed intervention does not have a substantial likelihood of ameliorating the targeted harm. To return to an aforementioned example—decreasing access to compute

[57] STEPHEN BREYER, REGULATION AND ITS REFORM 190 (1982) ("[R]egulatory failure sometimes means a failure to correctly match the tool to the problem at hand. Classical regulation may represent the wrong governmental response to the perceived market defect"). As a corollary, a dominant perspective—adopted in NAIAC's recommendation endorsing the NIST AI RMF—is that regulatory interventions should also be tied to level of risk. NAT'L A.I. ADVISORY COMM., NATIONAL ARTIFICIAL INTELLIGENCE ADVISORY COMMITTEE (NAIAC) YEAR 1 (2023), https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf [https://perma.cc/PG5V-9M63].

[58] Of course, if we're concerned about environmental externalities of energy-intensive operations, it's not clear why computing for AI models should be singled out. Much as is the case of banning LLMs to address bioterrorism risk, a tax on intensive computing to address environmental risks has its own alignment problem: if we care about internalizing the costs of climate externalities, there are strong reasons to prefer a general carbon tax, not one specific to AI computation.

[59] Jeanne Casusi, *What Is a Foundation Model? An Explainer for Non-Experts*, STANFORD INST. FOR HUMAN-CENTERED A.I. (May 10, 2023).

[60] We use the term "compute" to refer to the (often vast) computational resources required to train advanced AI models.

[61] Jai Vipra et al., *Computational Power and AI*, AINOW INSTITUTE (Sept. 27, 2023), https://ainowinstitute.org/publication/policy/compute-and-ai; Steve Lohr, *Universities and Tech Giants Back National Cloud Computing Project*, N.Y TIMES (June 30, 2022), https://www.nytimes.com/2020/06/30/technology/national-cloud-computing-project.html ("Fueling the increased government backing is the recognition that A.I. technology is essential to national security and economic competitiveness.").

may be mismatched to a goal of strengthening the AI innovation ecosystem because it restricts access to resources necessary for model development.

Regulatory Mismatch Arising from Unintended Consequences of Regulatory Objectives. Mismatch can also be more subtle and turn on nuances in the technical methods a regulation calls for. Recognizing the limits of conventional anonymization protocols,[62] some have turned to stronger measures, like differential privacy.[63] But whether a particular implementation of differential privacy achieves privacy goals depends on how practitioners configure the algorithm.[64] And absent any guidance about these settings, requirements to use differential privacy can reduce to mathematical window dressing.[65]

In algorithmic fairness, many companies have employed EEOC's 80% rule (that there is facial evidence of disparate impact if a protected group is selected at less than 80% of the rate of the majority group) as the quasi-regulatory objective to ensure algorithms are not biased.[66] Yet the 80% rule is merely guidance and neither encompasses the full thrust of antidiscrimination law[67] nor adheres to many other technical definitions of fairness.[68] In fact, the 80% rule is commonly implemented by discarding features that are highly correlated with protected attributes.[69] This could undermine underlying fairness values if the feature that is most predictive of job performance is discarded.[70] A credit algorithm that inaccurately scores individuals may not be more "fair."

Finally, mismatch can also occur when an intervention fails to address more systemic factors contributing to the harm. Returning to the bioweapons example, the restriction of

---

[62] Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* (Carnegie Mellon Univ., Data Privacy Working Paper No. 3, 2000), https://dataprivacylab.org/projects/identifiability/paper1.pdf.
[63] *Why the Census Bureau Chose Differential Privacy*, CENSUS.GOV (Mar. 27, 2023), https://www.census.gov/library/publications/2023/decennial/c2020br-03.html.
[64] Implementing differential privacy requires practitioners to set two numerical parameters, often referred to as *epsilon* and *delta*. The larger these parameters are, the less privacy is guaranteed. Setting these to large values is thus equivalent to not implementing differential privacy at all. See Kobbi Nissim, *Differential Privacy: A Concise Tutorial*, http://helper.ipam.ucla.edu/publications/pbd2018/pbd2018_14892.pdf.
[65] *See supra* Greenberg, note 56.
[66] *See* Christo Wilson et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening,* 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 666, 668 (2022); Elizabeth Anne Watkins et al., *The Four-Fifths Rule Is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness* 1, ARXIV (Feb. 19, 2022), https://arxiv.org/pdf/2202.09519.pdf.
[67] *Id*.
[68] Arvind Narayanan, *Tutorial: 21 fairness definitions and their politics*, YOUTUBE, https://www.youtube.com/watch?v=jIXIuYdnyyk; *see also* Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, 2018 ACM/IEEE INT'L WORKSHOP ON SOFTWARE FAIRNESS (2018), https://fairware.cs.umass.edu/papers/Verma.pdf; Dana Pessach & Erez Schmueli, *Algorithmic Fairness*, ARXIV (2020), https://arxiv.org/pdf/2001.09784.pdf.
[69] Feldman et al., *supra* note 56.
[70] *See* Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, PROCS. OF THE 8TH CONF. ON INNOVATIONS IN THEORETICAL COMPUT. SCI. 43 (2017).

large, open models may not fully achieve the underlying objective of minimizing the dissemination of bioweapons information because adversaries can rely on smaller or proprietary models to achieve the same end. Whether such restrictions are warranted to address biosecurity concerns therefore turns on the marginal impact of such restrictions on the diffusion of the relevant knowledge, and at what cost. The mainstream use of these models, and attempts to stress test them, has also brought heightened attention to insufficient lab safety protocols and other biosecurity vulnerabilities.[71] Although the diffusion of biosecurity risks could conceivably justify some restrictions on the diffusion of future versions of the most advanced LLMs, it is worth bearing in mind that such restrictions may ironically also undercut a broader societal goal of identifying regulatory gaps that, if closed, can reduce bioweapons risk.

Regulatory Proposals May Expose Value Conflicts ("Horizontal Misalignment"). Even if the regulatory value, objective, and behavior are aligned, a less recognized challenge is that values themselves conflict. Identifying bias, for instance, requires access to demographic data, but privacy's data minimization principle may make access to such demographic data challenging, posing a "privacy-bias tradeoff."[72] U.S. federal agencies, for instance, operate under a data minimization scheme established by the Privacy Act of 1974, which has posed serious challenges for conducting disparity assessments as mandated under the racial justice Executive Order: 21 of 25 agencies point to data challenges that impede equity impact assessments.[73]

Another example of horizontal misalignment lies in the tension between "international competitiveness" and "trustworthy AI." Seeking to win the geopolitical AI race[74] has generated legislative proposals to accelerate AI development, but such acceleration can be in tension with safeguards and protocols designed to slow development.[75] Proposals may

[71] *See, e.g.*, Service, *supra* note 11; Urbina et al., *supra* note 21; Vivek Wadhwa, *The Genetic Engineering Genie Is Out of the Bottle*, Foreign Policy (Sep. 11, 2020), https://foreignpolicy.com/2020/09/11/crispr-pandemic-gene-editing-virus/.

[72] Arushi Gupta et al., *The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government*, PROC. 2023 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 495 (2023).

[73] *Id.*; *see*, Exec. Order No. 13,985, 86 FR 7009 (7009-7013) (2021).

[74] For broader discussions explaining why winning the geopolitical competition, particularly with China and Russia, is critical for the United States, and its allies and partners, see Nat'l Sec. Comm'n on A.I., *2021 Final Report*, *supra* note 49; Special Competitive Stud. Project, *Mid-Decade Challenges to National Competitiveness* (2022), https://www.scsp.ai/wp-content/uploads/2022/09/SCSP-Mid-Decade-Challenges-to-National-Competitiveness.pdf.

[75] *Compare* Alexander C. Karp, *Our Oppenheimer Moment: The Creation of A.I. Weapons*, N.Y. TIMES (July 25, 2023) (advocating for an investment in the rapid development of AI weapon systems on par with the Manhattan Project), https://www.nytimes.com/2023/07/25/opinion/karp-palantir-artificial-intelligence.html, *with* Sigal Samuel, *The Case for Slowing Down AI*, VOX (Mar. 20, 2023, 7:58 AM) (calling for a slowdown in the development of advanced AI systems).

espouse values of transparent, privacy-preserving, non-discriminatory, explainable, *and* accurate AI as if they are all jointly achievable, but these horizontal misalignment issues mean that such values can quickly come into tension with one another in practice.

* * *

We have identified AI's regulatory alignment problem. Addressing it requires engaging with questions around compliance.[76] Do regulatory objectives further the chosen societal value? Does the behavior required for compliance comport with the objective? And how does one resolve the tension between values under full compliance? We now proceed to analyze these alignment problems for four of the most common AI regulatory proposals: disclosure, registration, licensing, and auditing.[77] Table 3 briefly describes common categories of AI regulatory proposals and identifies exemplars.

[76] *See* Giles, *supra* note 23.

[77] We selected these interventions because they are among the most commonly proposed, and therefore most relevant to current policy debates. We additionally note that while other interventions (like a compute-based tax) are not discussed here, the elements of our analysis could be extended to those interventions.

Intervention	Disclosures	Registration	Licensing	Audits
Description	Regulations requiring AI system developers or deployers to share information with the public at large about the system and any aspect of its performance, training data, design, or downstream applications.	Regulations requiring AI system developers or deployers to provide information about qualifying systems to government regulators, possibly accompanied by bans on use of unregistered models or penalties for non-registered use.	Regulations requiring entities like model developers to meet certain criteria prior to engaging in certain activities, like developing or deploying certain types of AI systems.	Regulations requiring verification by auditors that an AI system complies with relevant regulations, best practices, or standards.
Examples	Executive Order 13960;[78] Connecticut SB 1103[79]	EU AI Act;[80] Hawley-Blumenthal Framework[81]	Microsoft Blueprint;[82] Warren-Graham Bill (S. 2597);[83] Hawley-Blumenthal Framework[84]	NYC Bias Audit Law (Local Law 144)[85]
Key Design Features[86]				
Public information	Yes	No	Maybe	No
Government review or approval	No	Limited	Yes	Yes
Pre-market Requirement[87]	No	Yes	Yes	No

Table 3: Descriptions of four common AI regulatory proposals with examples.

For each form of regulation, we provide in Sections III through VI a description of the intervention, assess its technical and institutional feasibility, and connect it to the broader

[78] Exec. Order No. 13,960, 83 Fed. Reg. 65814 (Dec. 21, 2018).
[79] An Act Concerning Artificial Intelligence, Automated Decision-Making and Personal Data Privacy, S.B. No. 1103, Sess. Yr. 2023 (Ct. 2023).
[80] Press Release, European Parliament, *supra* note 17.
[82] Microsoft's blueprint for AI regulation calls for licensing of both large models and the data centers in which they are hosted. MICROSOFT, GOVERNING AI: A BLUEPRINT FOR THE FUTURE 20 (2023), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.
[85] Lindsay Stone, *NYC Issues Final Regulations for Automated Employment Decision Tools Law, Delays Enforcement to July 5, 2023*, JDSUPRA (Apr. 13, 2023), https://www.jdsupra.com/legalnews/nyc-issues-

themes of AI regulatory alignment problem. Before examining each regulatory intervention at length, we note several technical and institutional challenges that most, if not all, of these proposals will face.

First, the success of any regulatory scheme will depend critically on regulatory capacity, which itself will depend on the organization and presence of technical expertise within government agencies. For instance, a new AI super-regulator—something called for by proposals fitting within all four regulatory categories[88]—would run into major challenges given that a wide range of agencies already regulate AI products. A new agency would have to coordinate with or absorb regulatory authorities from (a) FDA's regulation of AI medical devices,[89] (b) HUD's oversight of algorithmic bias in housing,[90] (c) CFPB's regulation of AI used in consumer financial products,[91] (d) CPSC's protection of safety in consumer products,[92] (e) FTC's regulation of advertising claims and enforcement of

final-regulations-for-3612453/; N.Y.C. Dep't of Consumer & Worker Prot., *Notice of Adoption to Add Rules to Implement New Legislation Regarding Automated Employment Decision Tools* (July 5, 2023), https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/.

[83] Digital Consumer Protection Commission Act of 2023, S. 2597, 118th Cong.

[84] Blumenthal & Hawley, *supra* note 81.

[85] Lindsay Stone, *NYC Issues Final Regulations for Automated Employment Decision Tools Law, Delays Enforcement to July 5, 2023*, JDSUPRA (Apr. 13, 2023), https://www.jdsupra.com/legalnews/nyc-issues-final-regulations-for-3612453/; N.Y.C. Dep't of Consumer & Worker Prot., *Notice of Adoption to Add Rules to Implement New Legislation Regarding Automated Employment Decision Tools* (July 5, 2023), https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/.

[86] We describe what we understand to be the necessary design features of each category of regulation proposals. These classifications are *approximate*; specific regulatory proposals may have features that collapse distinctions between the categories.

[87] By this we mean that the regulatory intervention occurs before the AI model is released to the market (i.e., pre-market). For a non-AI example, think about FDA drug approvals that must occur before the drug is sold to consumers.

[88] *Infra* notes 265, 285.

[89] Eric Wu et al., *How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals*, 27 NATURE MED. 582 (2021).

[90] Press Release, U.S. Dep't of Just., Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising (June 21, 2022), https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known.

[91] Press Release, Consumer Fin. Prot. Bureau, CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms, https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/ [https://perma.cc/X2DM-MMUZ].

[92] CONSUMER PROD. SAFETY COMM'N, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN CONSUMER PRODUCTS (2021), https://www.cpsc.gov/s3fs-public/Artificial%20Intelligence%20and%20Machine%20Learning%20In%20Consumer%20Products.pdf.

consumer protections,[93] (f) DOT's oversight of self-driving cars,[94] (g) EEOC's examination of AI used in employment decisions,[95] (h) SEC's rulemaking around the use of AI by broker-dealers or investment advisors,[96] to name a few. Setting aside the hurdles legislation creating a new agency is likely to face, such a reorganization of government would be enormously complex. Evidence on the effectiveness of similar reorganizations has not been inspiring.[97] As James Q. Wilson said, "Presidents have taken to reorganizations the way… people take to fad diets—and with about the same results."[98]

Second, government agencies are in sore need of AI expertise, with fewer than 1% of new AI PhDs entering public service in 2022,[99] and the AI skills gap poses a serious threat to the effectiveness of any form of regulation. A new agency would likely confront the same issue. One potential approach to bridging this challenge—take inspiration from the former

[93] Michael Atleson, *Keep Your AI Claims in Check*, FED. TRADE COMM'N (Feb. 27, 2023), https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check [https://perma.cc/BQ9J-NAWJ]. Note that the FTC's jurisdiction over a nonprofit and other not-for-profit entity hinges on whether the entity actually operates for a profit. *See* Leonard L. Gordon, Nicholas M. Reiter, Allison B. Gottfried, Rebecca J. Lee, Imani T. Menard, George E. Constantine & Andrew L. Steinberg, *Comments and Challenges Welcome: FTC Proposes New Rule to Ban Non-Compete Agreements with Employees, Independent Contractors, and Volunteers*, VENABLE (Jan. 23, 2023), https://www.venable.com/insights/publications/2023/01/comments-and-challenges-welcome-ftc-proposes; Anna Lenhart, Senators Propose a Licensing Agency For AI and Other Digital Things, TECH POL'Y PRESS (Aug. 3, 2023), https://techpolicy.press/senators-propose-a-licensing-agency-for-ai-and-other-digital-things/ (explaining that because of the FTC's limited jurisdiction, "comprehensive privacy bills such as the American Data Privacy and Protection Act (ADPPA) often add the following language to the covered entity definition: *'is an organization not organized to carry on business for its own profit or that of its members*'").

[94] Press Release, U.S. Dep't of Transp., U.S. Department of Transportation Releases Automated Vehicles Comprehensive Plan (Jan. 11, 2021), https://www.transportation.gov/briefing-room/us-department-transportation-releases-automated-vehicles-comprehensive-plan [https://perma.cc/PQV3-ENLJ].

[95] The EEOC's AI and Algorithmic Fairness Initiative will lead to the issuance of technical assistance and guidance for the use of AI in employment contexts. *Artificial Intelligence and Algorithmic Fairness Initiative*, EEOC (last visited Sept. 19, 2023), https://www.eeoc.gov/ai. Private parties may obtain a "Notice of Right to Sue" from the EEOC if, after filing a charge with the EEOC, the EEOC is unable to finish its investigation. *Filing a Lawsuit*, EEOC (last visited Sept. 19, 2023), https://www.eeoc.gov/filing-lawsuit.

[96] Press Release, Sec. Exch. Comm'n, SEC Proposes New Requirements to Address Risks to Investors From Conflicts of Interest Associated With the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers (July 26, 2023), https://www.sec.gov/news/press-release/2023-140; Joshua Geffon & Aaron Ginsburg, *SEC Proposes Rules on the Use of AI by Registered Investment Advisers and Broker-Dealers*, JDSUPRA (Aug. 15, 2023), https://www.jdsupra.com/legalnews/sec-proposes-rules-on-the-use-of-ai-by-8228482/.

[97] *See, e.g.*, Jason Marisam, *Duplicative Delegations*, 63 ADMIN. L. REV. 181 (2011).

[98] JAMES Q. WILSON, BUREAUCRACY: WHAT GOVERNMENT AGENCIES DO AND WHY THEY DO IT (2d ed. 2000).

[99] NESTOR MASLEJ ET AL., THE AI INDEX 2023 ANNUAL REPORT 245 fig. 5.1.9 (2023), https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

Congressional Office of Technology Assessment[100] and create an executive-branch office to house AI policy experts, but upskilling, training, and recruitment in the civil service will remain important.[101]

Third, the proper distribution of liability between developers and deployers is not immediately apparent. To assign penalties, e.g., for failing to comply with regulation or to enable redress where AI systems cause harm or unintended consequences, requires assigning responsibility to organizations or individuals, establishing clear lines of liability for harm, and determining procedures for determining responsibility for violations. The development of foundation models raises questions about whether the developer or parties deploying the model downstream—including by integrating AI models into different applications—should be the primary party responsible for the impact AI systems have on users and the public at large. New York City's requirement that employers are responsible for audit requirements reflects a decision to place responsibility and liability on deployers.[102]

III. Disclosure

Disclosure has long been the favored child of American regulators and lawmakers, with hundreds of disclosure laws at both the state and federal level, spanning sectors as diverse

[100] *The Office of Technology Assessment*, U.S. GOV'T ACCOUNTABILITY OFF., https://www.gao.gov/products/103962 (last visited Aug. 31, 2023). The Office of Technology Assessment was a highly utilized, small agency that provided Congress analytical support on the impact of new and emerging technologies. Although suspended in 1995, it delivered—over its 23-year existence—"over 750 reports to Congress on a wide range of topics, including health, energy, defense, space, information technology, the environment, and many others; the vast majority of these reports were also made available to the public." Peter D. Blair, *Effective Science and Technology Assessment Advice for Congress: Comparing Options*, 48 SCI. & PUB. POL'Y 164, 167 (2021), https://doi.org/10.1093/scipol/scaa070. For proposals to bring back OTA, see Darrell M. West, *It Is Time to Restore the US Office of Technology Assessment*, BROOKINGS INST. (Feb. 10, 2021), https://www.brookings.edu/articles/it-is-time-to-restore-the-us-office-of-technology-assessment/.

[101] In some ways, the U.S. Digital Service operates in a similar vein but with a focus on directly assisting with technical implementation rather than policymaking. Of course, developing a shared resource for AI policy expertise can only have an impact on regulation to the extent that it is relied upon by other agencies. Designing such an office as an independent resource that's available for any who might seek it out could risk under-utilization, particularly if its staff become seen as lacking relevant policy domain expertise, while requirements for consultation or review pose a risk of resentment or creating cumbersome process that may slow down regulation of a field that is already moving so fast it is difficult for government to keep up, similar to the challenges that the Paperwork Reduction Act has posed to user-centered design research. *See* JENNIFER PAHLKA, RECODING AMERICA 140–143 (2023). This is an important institutional design challenge in its own right. Consultation could be mandated or this office could be vested with the power of publishing independent reports (along the lines of an Inspector General), which could increase transparency and improve alignment of agency actions with its recommendations, but may also create an atmosphere of mistrust that could result in agency staff keeping the office at arm's length, even if consultations were required.

[102] *See infra* notes 338–340 and accompanying text.

as securities, health and safety, and ethics.[103] Disclosure regulations typically require that entities provide certain information to the *public*, in contrast to registration schemes, which require that certain information be provided to the *government*. To its proponents, disclosure is regulation by light touch. In industries as fast evolving as AI, disclosure schemes can also identify potential harms and help to inform future public policy.[104]

It is thus unsurprising that there have been numerous proposals to enforce disclosure requirements on AI developers, deployers, and users. Disclosure proposals fall into three categories: institution-level, system-level, and prediction-level.

Institution-level disclosures target the procedures, practices, and organization of that institution. They provide information to consumers on the ways in which institutions utilize AI development and deployment across a *range* of applications. For instance, several proposals have called on governmental actors to create "algorithm registries" or "use case inventories", which disclose the different ways in which they use or rely on AI.[105] These disclosures provide insight into how particular agencies view AI and the types of activities they are willing (or unwilling) to automate.[106]

System-level disclosures target information about a specific AI system: for example, how it is used, how it was developed, and how it performs.[107] This can entail requirements to

[103] *See, e.g.*, Press Release, U.S. Sec. & Exch. Comm'n, SEC Adopts Rules on Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure by Public Companies (July 26, 2023), https://www.sec.gov/news/press-release/2023-139; National Bioengineered Food Disclosure Standard, 7 C.F.R. 66 (2019); Press Release, U.S. Sec. & Exch. Comm'n, Fact Sheet: ESG Disclosures for Investment Advisers and Investment Companies (May 25, 2022).

[104] *See, e.g.*, NAT'L TRANS. SAFETY BD., WE ARE ALL SAFER: LESSONS LEARNED AND LIVES SAVED (4th ed. 2006) (noting the thousands of safety regulations and advances that have derived from NTSB investigations and information-gathering activities), https://www.ntsb.gov/safety/safety-studies/Documents/SR0601.pdf; Justin Doubleday, *CISA Platform Helps Agencies Uncover More Than 1,000 Cyber Vulnerabilities*, FED. NEWS NETWORK (Aug. 25, 2023) (discussing CISA's vulnerability disclosure program and its resulting effect on agency security practices), https://federalnewsnetwork.com/cybersecurity/2023/08/cisa-platform-helps-agencies-uncover-more-than-1000-cyber-vulnerabilities/.

[105] *E.g.,* Exec. Order No. 13,960, *supra* note 78; Advancing American AI Act §7225, 40 U.S.C. §13301; *see also* Christie Lawrence, Isaac Cui & Daniel E. Ho, *The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies*, 2023 AAAI/ACM CONF. ON AI, ETHICS, & SOCIETY at 3 (2023) (discussing the implementation of AI registries across city, state, and federal agencies).

[106] Some scholars have noted, for instance, that excessive reliance on automation may call into question the very justifications for agency deference. Ryan Calo et al., *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L. J. 797 (2021).

[107] Rishi Bommasani et al., *Ecosystem Graphs: The Social Footprint of Foundation Models*, ARXIV (2023), https://arxiv.org/abs/2303.15772.

disclose if AI is used in a particular decision-making process,[108] the composition of training datasets,[109] whether collected data carries privacy or legal risks,[110] performance on public benchmarks,[111] and the potential for harmful use by downstream actors.[112]

Prediction-level disclosures, by contrast, target information about a specific prediction made by an AI system. Prediction-level disclosure requirements can encompass obligations to share when a particular prediction was AI generated,[113] the rationale behind a prediction,[114] what factors would alter the prediction generated,[115] or the level of certainty in the prediction.[116]

Not all disclosure requirements affecting AI will come from AI-centric regulation. "Rights" that the public receive explanations or specific information are scattered across American law, from the Due Process Clause of the 14th Amendment (disclosure of risk assessment score methodology in parole decisions) to the Equal Credit Opportunity Act (disclosure of loan denial reasoning through adverse action notification).[117] The question of how such laws interact with AI systems—across different legal contexts—has already been subject to litigation.[118]

A. Technical Feasibility: Disclosures May Require Information Not Possible To Collect

[108] *E.g.*, Off. Sci. & Tech. Policy, White House, *Blueprint for an AI Bill of Rights* (Oct. 4, 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights/ [hereinafter OSTP]; Fla. Stat. § 501.2041 (Fla. 2021); H.B. 20 § 120.052, 87th Leg., 2d Spec. Sess. (Tex. 2021).
[109] Khari Johnson, *Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI*, VENTUREBEAT (Sep. 28, 2020, 11:41 AM), https://venturebeat.com/ai/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/.
[110] Rishi Bommasani, Kevin Klyman, Daniel Zhang & Percy Liang, *Do Foundation Model Providers Comply with the Draft EU AI Act?*, CTR. FOR RSCH. ON FOUNDATION MODELS (June 15, 2023), https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.
[111] *Id*.
[112] *Id*.
[113] *E.g.*, AI Disclosure Act of 2023, H.R. 3831, 118th Cong. (2023).
[114] OSTP, *supra* note 108, at 6.
[115] Susanne Dandl & Christoph Molnar, *Counterfactual Explanations*, *in* INTERPRETABLE MACHINE LEARNING (2023), https://christophm.github.io/interpretable-ml-book/counterfactual.html.
[116] Charles Corbière et al., *Addressing Failure Prediction by Learning Model Confidence*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 32 (2019).
[117] Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 2045 (2021).
[118] *See, e.g.*, *Flores v. Stanford*, 18 CV 02468 (VB) (S.D.N.Y. Sep. 28, 2021) (requiring disclosure of information regarding how COMPAS scores are computed to plaintiff expert); *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016) (holding that a defendant's due process rights were not violated by the use of a risk assessment algorithm to inform a sentencing decision).

The technical feasibility of AI disclosure requirements will turn on whether the information demanded by the disclosure is capable of collection. Collectability focuses on the cognitive limits of developers operating with existing state-of-the-art AI auditing tools, the level of subjectivity implicated by different informational demands, and the technical difficulty in acquiring necessary information.[119]

A first question in assessing technical feasibility is how disclosure requirements may be affected by model size, training data, or prediction volume. Modern AI systems achieve large scales on all dimensions. Models like GPT-4 have trillions of parameters[120] and are trained on trillions of tokens.[121] When deployed as part of large platforms, they may be called upon to make millions of predictions per day, for tasks like search, ad-targeting, and content recommendations.[122] Laws which require developers to make disclosures on a per-datapoint or per-prediction level thus run the risk of being prohibitively costly. These include, for instance, requirements to share valuations of individual pieces of training data[123] or individualized explanations accompanying predictions.[124]

A second question is the extent to which a disclosure required by law is even possible to produce. Consider, for instance, the federal CIO Council's guidelines for algorithmic impact assessments.[125] The guidelines call for developers to "outline potential impacts or risks" of a project.[126] But are developers capable of assessing, *ex ante*, these impacts, which

---

[119] Sabri Eyuboglu et al., *Domino: Discovering Systematic Errors with Cross-Modal Embeddings*, 10 INT'L CONF. ON LEARNING REPRESENTATIONS, Apr. 2022 (noting that developers often work with high dimensional inputs, which make the deduction of higher-level observations regarding model behavior challenging).

[120] Reed Albergotti, *The secret history of Elon Musk, Sam Altman, and OpenAI*, SEMAFOR (Mar. 24, 2023, 11:09 AM), https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai.

[121] *See, e.g.*, Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, 36TH CONF. ON NEURAL INFORMATION PROCESSING SYSTEMS (2022) (1.4 trillion tokens); Hugo Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, META AI (July 18, 2023), https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/ (2 trillion tokens).

[122] Justin Burr, *9 Ways We Use AI in Our Products*, GOOGLE BLOG (Jan. 19, 2023), https://blog.google/technology/ai/9-ways-we-use-ai-in-our-products/.

[123] Amirata Ghorbani & James Zou, *Data Shapley: Equitable Valuation of Data for Machine Learning*, PROC. 36TH INT'L CONF. ON MACHINE LEARNING (2019).

[124] Kaminski & Urban, *supra* note 117, at 1980 (discussing the implications for complex AI systems of GDPR's creation of a "right to explanation" of automated decisions). Though–to the best of our knowledge–no proposals have been made requiring parameter-level analysis of models, such a policy would implicate similar concerns. *See* Clement Neo, *We Found An Neuron in GPT-2* (Feb. 11, 2023), https://clementneo.com/posts/2023/02/11/we-found-an-neuron.

[125] *Algorithmic Impact Assessment*, U.S. CHIEF INFO. OFFICERS COUNCIL (last visited Aug. 29, 2023), https://www.cio.gov/aia-eia-js/#/.

[126] *Id*.

ones are more or less likely, and how significant they will be?[127] Disclosure regimes which anchor too much on asking developers to prognosticate thus raise reliability concerns.

Another variant of this tension emerges when disclosures require developers to describe how a model produces predictions generally ("interpretability"), or why a specific prediction was provided ("explainability").[128] These types of disclosures sit on technically uncertain ground.[129] The challenge of understanding how AI models operate, or the reasons for a particular prediction, have inspired significant scholarly discussion.[130] The literature here has produced a number of approaches, which vary in technical implementation, cost, and the type of explanation generated.[131] There is little consensus on the *right* way to measure interpretability, with acknowledgement that interpretability depends on the type of data operated on, AI approaches, and explanation required.[132]

B. Institutional Feasibility: Effective Disclosures Require Agencies Have Technical Expertise And Capacity To Identify And Verify Relevant Information

Disclosure schemes are often appealing due to perceived low implementation costs.[133] Indeed, compared to the other interventions discussed in this Essay, disclosures are rather simple.[134] For example, regulators do not need to set up a scheme for defining and distributing licenses.[135] To implement a disclosure regime, regulators merely need to define

[127] As Yogi Berra said, "It is difficult to make predictions, especially about the future." Daniel P. Dickstein, Editorial: It's Difficult To Make Predictions, Especially About the Future: Risk Calculators Come of Age in Child Psychiatry, 60 J AM ACAD CHILD ADOLESC PSYCHIATRY 8, 950 (2020).
Securities law recognizes that forward-looking statements are inherently tentative and provides them safe harbor should they later prove inaccurate. 15 U.S.C. § 78u–5.
[128] OSTP, *supra* note 108; *Interpretability Versus Explainability*, AWS (2023), https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html
[129] Zachary C. Lipton, *The Mythos of Model Interpretability*, ACM QUEUE (July 17. 2018), https://queue.acm.org/detail.cfm?id=3241340.
[130] *Id*.
[131] Nadia Burkart et al., *A Survey on the Explainability of Supervised Machine Learning*, 70 JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH 245 (2021).
[132] Valerie Chen et al., Interpretable Machine Learning: Moving from Mythos to Diagnostics, 65 COMMUNICATIONS OF THE ACM 43 (2022).
[133] Daniel E. Ho, *Buyer Beware: With Mandated Disclosure, You Get What You Pay For*, DAILY J. (Feb. 22, 2017), https://dho.stanford.edu/wp-content/uploads/FinalCopy-3.pdf [hereinafter *Buyer Beware*].
[134] Cass R. Sunstein, *Informational Regulation and Informational Standing: Akins and Beyond*, 147 U. PA. L. REV. 613, 625 (1999).
[135] *See infra* Section V.B.

what information should be provided and the process by which it should be shared.[136] But implementing an *effective* disclosure regime requires more.

Several characteristics make disclosure requirements easier to implement than other regulatory interventions. First, disclosures are forgiving of regulatory inexpertise. Agencies can require the discloser (the regulated entity) to collect, store, and publicize disclosures. By shifting the burden of information collection and production to the regulated entity, agencies do not have to establish much in the way of regulatory infrastructure, relative to other regulatory proposals. In addition, regulators do not need to understand the finer points of LLM development to create disclosure requirements. They can simply require that developers share all information that is "relevant" or "material." Command-and-control style regulation (i.e., regulators provide strict instructions that regulated entities must follow to avoid penalties),[137] however, necessitates a finer grained understanding, because regulators are usually required to articulate specific, practically applicable standards.

Second, disclosures often require less consensus among stakeholders.[138] Consider the use of facial-recognition technologies (FRT) by police departments.[139] A law banning FRT would require broad consensus amongst lawmakers that the harms of FRT outweigh the benefits. In contrast, a law mandating that police departments disclose FRT usage only requires consensus on the notion that transparency about FRT usage is relevant to the public.

That said, despite disclosure's theoretical appeal and widespread adoption, there remains significant debate as to the conditions that make disclosure effective. For instance, how

[136] A disclosure-only regime without a complementary oversight mechanism may have reduced effectiveness. Hans B. Christensen, Luzi Hail & Christian Leuz, *Mandatory IFRS reporting and changes in enforcement*, 56 J. ACCT. & ECON. 147 (finding that the benefits of disclosure were concentrated in locations with concurrent increases in regulatory enforcement); Colleen Honigsberg, *Hedge Fund Regulation and Fund Governance: Evidence on the Effects of Mandatory Disclosure Rules*, 57 J. ACCT. RSCH. 845 (2019) (finding benefits of disclosure without regulatory enforcement with sophisticated consumers).

[137] For discussion of command-and-control regulation in other sectors, *see, e.g.,* Hannah L. Baldwin, *Clearing the Air: How an Effective Transparency Policy Can Help the U.S. Meet its Paris Agreement Promise*, 35 J.L. & COM. 79 (2016); Dan Farber, *Continuity and Transformation in Environmental Regulation*, 10 ARIZ. J. ENV'T L & POL'Y 1 (2019); Kristin Madison, *Health Care Quality Reporting: A Failed Form of Mandated Disclosure?*, 13 IND. HEALTH L. REV. 310 (2016); Vincent R. Johsnon, *Nanotechnology, Environmental Risks, and Regulatory Options*, 121 PENN. ST. L. REV. 471 (2016).

[138] Disclosures often have a broader political coalition than other interventions. Omri Ben-Shahar et al., *More Than You Wanted to Know: The Failure of Mandated Disclosure* 5 (2014).

[139] Clare Garvie et al., *The Perpetual Line-up: Unregulated Police Face Recognition in America* (2016), https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%201 21616.pdf.

much of disclosure's success in the securities regime can be attributed to disclosure, and how much is due to unique aspects of the securities ecosystem?[140] Skeptics would argue that securities disclosure thrives within a comparatively robust private enforcement regime, in which well-resourced plaintiffs (e.g., shareholders) can bring high-value claims for omissions and misstatements. Securities disclosure also benefits from a robust network of intermediaries (e.g., securities analysts), which explicitly and implicitly translate complex technical disclosures into informational signals (e.g., share prices) that ordinary consumers can understand.[141] The fluidity of the securities market—in which purchasers can exercise an extraordinary amount of choice—makes disclosures actionable for consumers.

Additionally, some scholars have argued that crafting effective disclosure may in fact be neither cheap nor easy.[142] Agencies need to know what information to ask for, which can be difficult without AI expertise or prior knowledge of, or transparency into, the AI systems companies are developing. Regulated entities may protest that disclosures implicate significant trade secrecy or privacy concerns, especially when disclosures pertain to proprietary approaches, user behavior, or data.[143] However, increased secrecy could compound information asymmetries between companies and the public on AI usage.

Ensuring that disclosures are accurate often requires regulators to fall back on the traditional command-and-control style interventions and invest significant resources to verify information. The targets of disclosure laws often spend huge amounts of time and money on information reporting.[144] And financial disclosure requirements have led hedge funds to change their internal governance, which increased the accuracy of mandatory reporting.[145] However, identifying misstatements and omissions requires auditing personnel or internal whistleblowers. Private enforcement requires investment in personnel to operate tribunals and adjudicate claims. Skeptics could argue that targeted entities could spend these resources on making their offered product safer or more effective.[146] Of course,

---

[140] Paula J. Dalley, *The Use and Misuse of Disclosure as a Regulatory System*, 34 FLA. ST. U. L. REV. (2007).
[141] *See generally* Eugene F. Fama, *Efficient Capital Markets: A Review of Theory and Empirical Work*, 25 J. FIN. 383 (1970).
[142] *See* OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2014).
[143] Daphne Keller, *User Privacy vs. Platform Transparency: The Conflicts are Real and We Need to Talk About Them*, Ctr. For Internet and Soc. (Apr. 6, 2022), https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0; Tyler Trew, *Ethical Obligations in Technology Assisted Review*, ABA PRACTICE POINTS (Dec. 7, 2020), https://www.americanbar.org/groups/litigation/committees/professional-liability/practice/2020/ethical-obligations-in-technology-assisted-review/
[144] Sunstein, *supra* note 134.
[145] Colleen Honigsberg, Hedge Fund Regulation and Fund Governance: Evidence on the Effects of Mandatory Disclosure Rules, 57 J. ACCT. RSCH. 4, 845 (2019).
[146] *Buyer Beware*, *supra* note 133.

verifying disclosed information is not necessarily required in disclosure regimes, but this begs the question whether unreliable disclosures are useful.

An example of disclosure's "practical" challenges is provided by the federal government's experience implementing an AI registry. The 2020 Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (EO 13,960) required agencies to create and publicly disclose an inventory of AI use cases (i.e., an AI registry).[147] On the one hand, it might have seemed easy to simply require this disclosure. Yet the precursor to that mandate illustrates the challenge: it took a team of some 30 students at Stanford over a year to complete the inventory, with AI use cases spread across hundreds of agencies and officials, each with dramatically varying definitions and understandings of AI.[148]

The federal government's own effort at documenting use cases resulted in dramatic inconsistencies across agencies, with, at best, half of agencies with demonstrable AI use cases making public an inventory.[149] Customs and Border Protection (CBP), for instance, refused to classify its facial biometric scanning as AI.[150] Conducting such an inventory requires expertise, personnel, rules for defining AI and exemptions, and adjudication of boundary issues. The same challenges haunted New York City's Automated Decision Systems Task Force, with city officials expressing concern that regulations "would apply to every calculator and Excel document."[151]

The balance of disclosures' costs and benefits in other fields offers guiding principles when thinking about its application for AI.[152] The literature suggests that disclosures are most

[147] Exec. Order No. 13,960, 83 Fed. Reg. 65814 (Dec. 21, 2018).
[148] DAVID FREEMAN ENGSTROM ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 13–14 (Feb. 2020) (report submitted to the Admin. Conf. U.S.), https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf. The challenges faced today have been documented by Lawrence, Cui & Ho, *supra* note 105, and subsequent reporting. *E.g.*, Rebecca Heilweil and Madison Adler, OMB Acknowledges Issues with Process for Inventorying AI use cases, FEDSCOOP (Aug. 16, 2023), https://fedscoop.com/omb-acknowledges-issues-with-process-for-inventorying-ai-use-cases/; Rebecca Heilweil and Madison Adler, Agency AI inventories expected to get attention from House oversight subcommittee, FEDSCOOP (Sept. 13, 2023), https://fedscoop.com/agency-ai-inventories-expected-to-get-attention-from-house-oversight-subcommittee/.
[149] Lawrence, Cui & Ho, *supra* note 105.
[150] *Compare* ENGSTROM ET AL., *supra* note 148, at 31–32 (discussing CBP's extensive use of facial recognition software which utilized deep learning and other machine learning methods), *with Artificial Intelligence Use Case Inventory*, Dep't Homeland Sec., https://www.dhs.gov/data/AI_inventory (last visited Sep. 14, 2023) (listing no facial recognition or other biometric identification AI use cases for CBP).
[151] Albert Fox Cahn, *The First Effort to Regulate AI Was a Spectacular Failure*, FAST CO. (Nov. 26, 2019), https://www.fastcompany.com/90436012/the-first-effort-to-regulate-ai-was-a-spectacular-failure.
[152] ARCHON FUNG, MARY GRAHAM & DAVID WEIL, FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2007).

effective when they meet three criteria: understandability, actionability, and verifiability.[153] Disclosures filled with jargon or excessive detail will overwhelm ordinary consumers, who often lack technical expertise necessary to understand the disclosures. In addition, because individuals derive value from comparing the information contained in different disclosures, ensuring that disclosers follow standards with regards to terminology and form are essential to ensure understandability.

In the context of AI, regulators have two paths for addressing understandability. First, they can mandate that disclosures are structured in forms that are comprehensible to a wide range of stakeholders. For ordinary consumers, this could involve requirements that disclosures adhere to a plain-language standard.[154] Regulators could also consider experimenting with more interactive forms of disclosure, which tailor the information offered to an individual (e.g., through APIs).[155] Regulators must also be wary of disclosure fatigue. Disclosures which are too frequent—like California's Prop. 65 "carcinogenic" warning[156] or the EU's website cookie notifications[157]—are often ignored by consumers.

Alternatively, regulators can implement disclosure in settings where an information intermediary[158] is present. In medical contexts for instance, regulators can rely on doctors to parse disclosures associated with medical machine learning systems, and accurately communicate potential risks and benefits to patients.[159] While this presumes some level of expertise on the part of the intermediary, there are indications that specialized disciplines like law are increasingly viewing familiarity with AI as a skill essential to the profession.[160]

The second criterion of effective disclosure, actionability, pertains to the disclosure recipient's level of agency. If recipients have no opportunity to apply the information to decision-making, then disclosures will be less impactful. Ideally, consumers would have

[153] Omri Ben-Shahar and Carl E. Schneider, *The Failure of Mandatory Disclosure*, 159 U. PA. L. REV 647 (2011).
[154] OSTP, *supra* note 108.
[155] *SEC Disclosure API Available*, SEC (Sep. 8, 2021), https://www.sec.gov/structureddata/announcement/osd-announcement-090821-sec-disclosure-data-api.
[156] *See* Lisa A. Robinson et al., *Efficient Warnings, Not "Wolf or Puppy" Warnings*, *in* THE FUTURE OF RISK MANAGEMENT 227 (Howard Kunreuther et al. eds., 2019).
[157] Charlie Warzel, *Slouching Towards 'Accept All Cookies'*, ATLANTIC (Sep. 12, 2023), https://www.theatlantic.com/technology/archive/2023/09/personal-data-digital-privacy-value-choices-rights/675183/.
[158] For a definition of information intermediary, see, e.g., *Information Intermediary*, OXFORD REFERENCE, https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100003398 (last visited Sept. 15, 2023).
[159] This is analogous to how doctors already parse and communicate risks for different procedures and medications.
[160] Julia Brickell et al., *AI Pursuit of Justice & Questions Lawyers Should Ask*, BLOOMBERG LAW (April 2022), https://www.bloomberglaw.com/external/document/X3T91GR8000000/tech-telecom-professional-perspective-ai-pursuit-of-justice-ques.

lots of options for products.[161] To put it more concretely: disclosures about fuel efficiency are more significant when an individual is picking between two family-friendly minivans. But if that individual is picking between a two-seater convertible and a minivan, fuel efficiency is not likely to be a decisive factor.

In the context of AI, regulators should identify decision points that AI users face, and design disclosures which inform the choices available at these junctures. Two specific decision points are worth highlighting. The first is when buyers choose to purchase an AI system.[162] Disclosures regarding the capabilities of offered systems could influence their eventual decision amongst different vendors. The second decision point is when individuals must decide whether to adhere to the recommendation or forecast of an AI system. For instance, where a doctor must decide whether to follow a diagnostic algorithm's prediction or conduct additional tests,[163] information about the certainty of the prediction or the reliability of the underlying system can shape the doctor's reliance on the diagnostic algorithm.[164]

Importantly, this suggests that disclosures providing measures of performance and information on how AI systems were evaluated may be most effective. For instance, developers of medical AI systems could be required to report group-level performance statistics.[165] The advantages of such disclosures over ones targeted at system construction (e.g., what training data was used) is two-fold. First, because AI researchers are still understanding how aspects of system design—such as model architecture or training data choices—influence model behavior, simply knowing that a model was trained on data from a particular source may be unhelpful.[166] Second, evaluation disclosures are better suited for providing information relevant to the decision criteria that disclosure recipients will rely

[161] Arguably, mandating disclosure can change the behavior of the entity required to disclose through, for example, public shaming.
[162] Deloitte, *AI Procurement in a Box: AI Government Procurement Guidelines* (June 2020), https://www2.deloitte.com/content/dam/Deloitte/global/Documents/About-Deloitte/gx-wef-ai-government-procurement-guidelines-2020.pdf.
[163] W Nicholson Price II et al., *Potential Liability for Physicians Using Artificial Intelligence*, 322 JAMA 1765 (2019).
[164] *See* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Human in the Loop*, 76 VAND. L. REV. 429 (2023).
[165] Solon Barocas et al., *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*, ARXIV (Dec. 1, 2021), https://arxiv.org/abs/2103.06076.
[166] A stark example is offered by Meta's efforts to train a LLM on "a high-quality and highly curated" collection of scientific publications. Ross Taylor et al., *Galactica: A Large Language Model for Science*, ARXIV (Nov. 16, 2022), https://arxiv.org/abs/2211.09085. Despite Meta's use of curated scientific text, the model nonetheless exhibited similar tendencies to LLMs trained on unfiltered web corpora. Aaron J. Snoswell & Jean Burgess, *The Galactica AI Model Was Trained on Scientific Knowledge – But It Spat Out Alarmingly Plausible Nonsense*, THE CONVERSATION (Nov. 29, 2022), https://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nonsense-195445 (describing the model's generation of false scientific information and other biased/toxic outputs).

on. When choosing which model to purchase, or whether to follow a model's prediction, disclosures about a model's expected accuracy, probability of error, or confidence are more informative.

Finally, disclosures should be verifiable.[167] One of the monikers for AI disclosures is that they are like *nutrition* score cards for AI.[168] Yet that analogy misses a central weakness of food law's disclosures: while the system strives for extreme transparency, such transparency may be a false promise when few actors are able to verify the information disclosed by manufacturers. [169] Due to the decline of random sampling (audits) of food labels by FDA, dwindling enforcement efforts, and compromises in disclosure rules,[170] the Government Accountability Office concluded, "the accuracy of the information provided on about 500,000 labels will depend largely on the food industry."[171] AI use case inventories, model cards, data sheets, and disclosures may fail if they cannot be verified.[172]

C. Disclosure's Tensions: Disclosures May Be Self-Defeating, Ineffective, Or Disproportionally Burden Regulated Entities

Where does that leave us? To assess the efficacy of disclosure, we assess the ability of disclosure to mitigate *information deficits* in AI. As noted above, there is limited information about the harms, and magnitude of those harms, caused by using AI systems. Disclosure requirements in the form of mandated or voluntary reporting of adverse events may reduce those deficits by promoting public transparency.[173] Of course, not all adverse events are equally ascertainable. But much like the FDA maintains a public system of adverse events in the drug context,[174] we would begin to know a lot more about AI's relative harms through such a reporting system.

---

[167] One question is whether consumers need the ability to verify or whether a third-party, such as a government entity or auditor, could perform this function.
[168] *See, e.g.*, Sara Gerke, *'Nutrition Facts Labels' for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards*, 91 Geo. Wash. L. R. 79 (2023).
[169] Lisa Heinzerling, *The Varieties and Limits of Transparency in US Food Law*, 70 FOOD & DRUG L. J. 11, 16 (2015) ("The USDA long ago abandoned any effort to conduct random sampling of food products . . . . Almost twenty years ago, FDA likewise abandoned any effort to conduct random sampling and analysis.").
[170] *Id.* ("The USDA long ago abandoned any effort to conduct random sampling of food products . . . . Almost twenty years ago, FDA likewise abandoned any effort to conduct random sampling and analysis.").
[171] U.S. Gov't Accountability Office, GAO/RCED-95-19, Nutrition Labeling: FDA and USDA Need a Coordinated Assessment of Food Label Accuracy 8 (1994).
[172] *See* Rishi Bommasani et al., *Ecosystem Graphs: The Social Footprint of Foundation Models*, ARXIV (Mar. 28, 2023), https://arxiv.org/abs/2303.15772 (proposing one approach to tracking and verifying the provenance of information relating to a foundation model, its training data, and its downstream applications).
[173] For a broader discussion of adverse event reporting, see *infra* Section IV.
[174] *FDA Adverse Event Reporting System (FAERS) Public Dashboard*, U.S. FOOD & DRUG ADMIN. (Oct. 22, 2021), https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard.

Yet disclosure is not without its drawbacks. First, disclosure itself may be self-defeating and create risks by increasing compliance costs and providing more information about AI systems. Increased transparency about AI systems could enable adversaries to manipulate such systems or learn sensitive information contained within training data, reflecting tensions between transparency promoted by disclosure policies and values of effectiveness or privacy.[175] For instance, platforms have long resisted calls for increased transparency about content moderation algorithms, arguing that such transparency would enable individuals to circumvent these models.[176] From a geopolitical perspective, transparency could hurt American competitiveness, by forcing model developers to publish technical details about their systems. Transparency could also harm a firm's competitive advantage. For example, scholars have identified how the Freedom of Information Act (FOIA)—originally envisioned as a tool for promoting public transparency—has largely been co-opted as a form of legalized corporate espionage.[177]

Also, it is important to recognize that the need to comply with disclosure laws could negatively affect how AI models are developed and maintained. For instance, because higher-capacity AI models tend to be less interpretable, a disclosure requirement that effectively mandates a certain level of interpretability could force developers to choose less accurate (but more explainable) models over more accurate (but less explainable) ones.[178] Similarly, disclosure laws requiring substantial manual processes for each model release would slow the rate at which developers can update deployed models. Because updates to models often address important deficiencies in performance, burdening developers' update speed could prevent performance gaps from being quickly addressed.[179]

Second, the information provided by disclosures may fail to have the intended effect. Sometimes disclosure requirements can actually *worsen* individual decision-making: in several studies, disclosure of conflicts-of-interest by an advisor led to worse decisions by

---

[175] Keller, *supra* note 143; Laura Edelson, *Platform Transparency Legislation: The Whos, Whats and Hows*, LAWFARE (April 29, 2022), https://www.lawfaremedia.org/article/platform-transparency-legislation-whos-whats-and-hows.

[176] Twitter, *Trust and Safety Models*, GITHUB, https://github.com/twitter/the-algorithm/tree/main/trust_and_safety_models (last visited Oct. 5, 2023).

[177] Margaret B. Kwoka, *FOIA, Inc.*, 65 DUKE L.J. 1361 (2016); April Klein et al., *Seeking Out Non-Public Information: Sell-Side Analysts and the Freedom of Information Act*, 95 ACCT. REV. 233 (showing how financial analysts use FOIA requests to improve stock predictions at healthcare companies).

[178] Anna Nesvijevskaia et al., *The accuracy versus interpretability trade-off in fraud detection model*, 3 DATA & POLICY e12 (2021). *But see*, Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NAT. MACH. INTELL. 206 (2019).

[179] Shreya Shankar et al., *Operationalizing Machine Learning: An Interview Study*, ARXIV (2022), https://arxiv.org/abs/2209.09125

the advisee, because "advisors [felt] comfortable giving more biased advice," while "advisees [felt] more uncomfortable in turning down advice (e.g., it signals distrust of the advisor's motives").[180] Disclosures are used to build public trust in a particular market or institution, but incomplete or incorrect disclosures may provide false assurance (e.g., the CBP's aforementioned neglect to disclose its heavily-used facial-recognition system in its AI use case inventory[181]). Such omissions, when revealed, may undermine trust in the broader disclosure system's efficacy. If not carefully designed, disclosure requirements may worsen the risks they set out to address.

Third, non-AI-specific regulation may provide a more effective solution than mandatory disclosures. For example, while disclosure has been proposed as a tool for addressing environmental harms,[182] transparency requirements are a meager substitute for more direct, impactful environmental interventions, such as investments in cleaner sources of power, improved grid infrastructure, or carbon-based taxes. Disclosures may provide a politically palatable solution without providing for direct action or recourse.

Fourth, in many cases, disclosure functions more as an audit or registration requirement. A disclosure requirement under which developers must compute fairness metrics on model performance for certain populations of data can function as an internal bias audit. A requirement that audits be conducted and shared can function as a disclosure. And disclosure requirements which require extraordinary transactional costs effectively operate as a licensing scheme or ban, limiting the number of entities capable of developing AI.

Last, disclosure may have disproportionate distributive impacts, advantaging well-resourced incumbents in the AI industry. Smaller developers or deployers may have more difficulty complying with disclosure requirements.[183] In the case of Executive Order 13,960, large government agencies (i.e., those subject to the Chief Financial Officers Act) were more easily able to comply with use case inventory requirements, while smaller agencies struggled.[184] In addition, disclosure may facilitate anti-competitive behavior. A study of gasoline price disclosures, for instance, showed that mandated disclosures

[180] *See* George Loewenstein et al., *The Limits of Transparency: Pitfalls and Potential of Disclosing Conflicts of Interest*, 101 AMER. ECON. REV. 423 (2011).
[181] *See supra* note 150.
[182] Bommasani, Klyman, Zhang & Liang, *supra* note 110.
[183] To balance the benefits of disclosure with the costs of providing this information, securities laws provide tiers of disclosure, with "smaller reporting companies" subject to fewer requirements. *Smaller Reporting Companies*, U.S. SEC. & EXCH. COMM'N (Apr. 6, 2023), https://www.sec.gov/education/smallbusiness/goingpublic/SRC.
[184] Rebecca Heilweil & Madison Alder, *The Government Is Struggling to Track Its AI. And That's a Problem*, FEDSCOOP (Aug. 3, 2023), https://fedscoop.com/the-government-is-struggling-to-track-its-ai-and-©-a-problem/.

softened competition, particularly in lower-income areas, as operators could coordinate more easily.[185]

IV. Registration

In contrast to disclosure—which promotes public transparency about AI systems and their behavior—AI registration proposals primarily seek to facilitate government awareness and oversight of technological capabilities, individual AI applications, and risks related to the use of AI. Registration[186] is often employed to increase safety, protect consumers, and strengthen national security.[187] Registration requires providing to the government, often through filing documentation and the payment of administrative fees,[188] information about specified activities, entities, individuals, or holdings, including facilities or other assets. Changes to the information provided often must be updated, either within specified time periods, after material changes, or both.[189] Registration is generally required *before* an entity or individual can engage in the specified activity (e.g., selling securities).[190] Failure to register can result in sanctions—either fines or other penalties.[191] For example, under the under the Foreign Agent Registration Act (FARA), individuals who agree to act as agents of foreign principals must register with the Department of Justice within ten days

[185] Fernando Luco, *Who Benefits from Information Disclosure? The Case of Retail Gasoline*, 11 AMER. ECON. J.: MICROECONOMICS 277 (2019).
[186] Regulators do not clearly distinguish between registration and licensing. Thus, there are some "registration" regimes that function more like licensing regimes. Here, we focus only on regimes that function as registration regimes.
[187] *See, e.g.*, *Registration and Listing*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/industry/fda-basics-industry/registration-and-listing (last visited Sep. 14, 2023); Dan Greene et al., *The Danger of Invisible Biolabs Across the U.S.*, TIME (Aug. 31, 2023, 8:40 AM), https://time.com/6309643/invisible-biolabs/; https://consteril.com/biosafety-levels-difference/; *Who Must Register*, U.S. DEPARTMENT OF STATE DIRECTORATE OF DEFENSE TRADE CONTROLS (last visited Oct. 6, 2023), www.pmddtc.state.gov/ddtc_public/ddtc_public?id=ddtc_kb_article_page&sys_id=7110b98edbb8d30044f9ff621f96192d; WHITNEY K. NOVAK, CONG. RSCH. SERV., IF11439, FOREIGN AGENTS REGISTRATION ACT (FARA): A LEGAL OVERVIEW (2023); MARC LABONTE, CONG. RSCH. SERV., R44918, WHO REGULATES WHOM? AN OVERVIEW OF THE U.S. FINANCIAL REGULATORY FRAMEWORK (2020); Paul J. Larkin, *Public Choice Theory and Occupational Licensing*, 38 HARV. J. L. PUB. POL. 209 (2015); Carolyn Cox et al, THE COSTS AND BENEFITS OF OCCUPATIONAL REGULATION, U.S. FEDERAL TRADE COMMISSION 49 (1990) https://www.ftc.gov/system/files/documents/reports/costs-benefits-occupational-regulation/cox_foster_-_occupational_licensing.pdf.
[188] Fees are often used to cover administrative costs. *See, e.g.*, Occupational licensing, *supra* note 168 (https://www.ftc.gov/system/files/documents/reports/costs-benefits-occupational-regulation/cox_foster_-_occupational_licensing.pdf at 49); https://www.fda.gov/medical-devices/how-study-and-market-your-device/device-registration-and-listing.
[189] For example, the Foreign Agent Registration Act requires foreign agents file supplemental statements within six months of the initial filing. 22 U.S.C. § 612(b). Foreign agents must also provide notice to the Department of Justice of any changes to the information provided, with the Attorney General authorized to require supplemental filings as determined necessary. 22 U.S.C. § 612(b); 28 C.F.R. § 5.203; Foreign Agents Registration Act, 22 U.S.C. § 611–621 (1938).
[190] *See e.g.,* Securities Act of 1933, 15 U.S.C. § 77e.
[191] *Id*.

of the agreement and may not engage in a covered activity without registering.[192] Willful violations are punishable by imprisonment and fines, or both.[193]

Registration requirements that apply to entities or individuals engaged in certain activities often also require registering details about activities. For example, the Securities and Exchange Act requires registration of companies selling securities and classes of securities.[194] Registered foreign agents under FARA must also provide the Department of Justice with copies of disseminated informational materials that clearly include a "conspicuous statement" that they are distributed by an agent, with such materials accessible through the Department to the public.[195] Thus, registration may include elements of mandatory disclosures.

Registration may also impose additional requirements, such as compliance with agency rules or additional oversight. Registration regimes may be tiered with these additional requirements applying to only a subset of the covered entities or activities. For example, the FDA requires medical device manufacturers to register and "list" any medical devices in commercial distribution before they can sell listed devices.[196] Devices in the lowest risk class are subjected to post-market "general controls." However, devices that pose a higher risk of injury face "special controls" like post-market surveillance, pre-market approval, and pre-approval manufacturing inspections.[197] However, all registered medical device manufacturers and importers must report to the FDA certain adverse incidents like deaths, serious injuries, and malfunctions. The information is made publicly available in the FDA Adverse Events Reporting System (FAERS) database, which the FDA uses to support post-market surveillance by identifying, monitoring, and analyzing risks.[198]

---

[192] 22 U.S.C. § 611 et seq; Foreign Agents Registration Act, 22 U.S.C. § 611–621 (1938).
[193] Punishable with up to five years imprisonment, a $250,000 fine, or both. *Foreign Agents Registration Act — Frequently Asked Questions*, U.S. DEP'T OF JUST., https://www.justice.gov/nsd-fara/frequently-asked-questions.https://www.justice.gov/nsd-fara/frequently-asked-questions (last visited Oct. 4, 2023).s)
[194] Securities Act of 1933, 15 U.S.C. §§ 77a–77mm (1934); MARCO LABONTE, CONG. RSCH. SERV., R44918, WHO REGULATES WHOM? AN OVERVIEW OF THE U.S. FINANCIAL REGULATORY FRAMEWORK 17 (2020).
[195] The materials must be filed within 48 hours of dissemination. 22 U.S.C § 614.
[196] *Medical Device Listing*, MED. RISK MGMT., https://www.medical-risk.com/en/regulatory-services/us-agent-services/medical-device-listing (last visited Oct. 5, 2023).
[197] Jeffrey K. Shapiro, *Substantial Equivalence Premarket Review: the Right Approach for Most Medical Devices*, 69 FOOD & DRUG L.J. 365, 372 (2014); Colleen Smith, *Scouting For Approval: Lessons on Medical Device Regulation in an Era of Crowdfunding from Scanadu's "Scout",* 70 FOOD & DRUG L.J., 209, 220 (2015); Sarah Lykken, *We Really Need to Talk: Adapting FDA Processes to Rapid Change*, 68 FOOD & DRUG L.J. 357, 374 (2013).
[198] *Mandatory Reporting Requirements: Manufacturers, Importers and Device User Facilities*, U.S. FOOD AND DRUG ADMINISTRATION (last updated May 22, 2020), https://www.fda.gov/medical-devices/postmarket-requirements-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities; *Questions and Answers on FDA's Adverse Event Reporting System (FAERS)*, U.S.

Current proposals suggest two approaches to AI registration. The first approach calls for registration of only sufficiently advanced models. Proponents argue that such models potentially create dire risks and thus merit distinctive regulatory attention.[199] What qualifies as advanced is subject to debate, with proffered criteria including whether the model has dangerous capabilities,[200] is a "sophisticated" general-purpose model,[201] is merely more advanced than the current generation of large models (e.g., GPT-4),[202] meets a size or FLOPs[203] threshold,[204] or achieves certain scores on public benchmarks (e.g., achieves an SAT score of at least 1300).[205] The second approach calls for developers to register models used in certain "high risk" domains. For instance, the EU AI Act would require registration in an EU database of AI models used in eight specific areas (e.g., biometric identification, law enforcement).[206] The rationale espoused for registration here can be found in the name itself—by their very nature, or at least as argued, model deficiencies are more likely to result in harmful consequences to individuals in these areas. Some proposals combine aspects of both approaches. For example, Senators Richard Blumenthal and Josh Hawley's recent "Bipartisan Framework for U.S. AI Act" would require companies that develop "sophisticated general-purpose A.I. models" or "models used in high-risk situations" to register "with an independent oversight body" and participate in incident reporting programs.

Though proposals disagree on *which* models should be registered, they largely agree on what registration should entail. Most call for government to create and maintain a database

---

FOOD AND DRUG ADMINISTRATION (last updated June 4, 2018), https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers; *Shapiro*, *supra* note 197; Suranjan De, *FDA Adverse Event Reporting System (FAERS) Reporting and Review*, U.S. FOOD AND DRUG ADMINISTRATION 5 (last visited Oct. 6. 2023), https://www.fda.gov/media/165667/download.

[199] *E.g.*, Gillian Hadfield et al., *It's Time to Create a National Registry for Large AI Models*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 12, 2023), https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180; Rishi Iyengar, *OpenAI's CEO Goes on a Diplomatic Charm Offensive*, FOREIGN POLICY (June 20, 2023), https://foreignpolicy.com/2023/06/20/openai-ceo-diplomacy-artificial-intelligence/; Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, ARXIV (July 11, 2023), https://arxiv.org/abs/2307.03718.

[200] Anderljung et al., *supra* note 199, at 2 ("[W]e focus on what we term "frontier AI" models—highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety.").

[201] Blumenthal & Hawley, *supra* note 81; Press Release, European Parliament, *supra* note 17.

[202] Hadfield et al., *supra* note 199.

[203] S*ee infra* notes 217–219 and accompanying text (discussion on technical feasibility of FLOPs proposals).

[204] Hadfield et al., *supra* note 199.

[205] *A Responsible AI Act*, CENTER FOR AI POLICY, https://www.aipolicy.us/work (last visited Oct. 6, 2023).

[206] Press Release, European Parliament, *supra* note 17.

listing developers and providing information about covered AI models.[207] Information should allow for transparency into the design and structure of these models, and thus contain details on architecture, size, training processes, and training data.[208] Proposals also agree that operating or using unregistered models should be banned.[209] Several proposals call for coupling this registration with additional mechanisms, like licensing, incident reporting, or novel oversight bodies. [210]

A. Technical Feasibility: Registration Criteria May Not Track Risk

AI registration is only technically feasible if it is possible to identify which systems meet registration criteria. Without a clear understanding of which systems should be registered and what qualities of those systems necessitate registration, regulators will be unable to effectively police non-compliance. Moreover, if criteria are highly subjective, then enforcement of registration will be inconsistent, frustrating regulatory goals.

A version of this concern emerges in proposals which premise registration on whether models possess capabilities which might lead to catastrophic harms.[211] But reducing this inquiry into an objective, measurable standard that a regulator can implement is far from clear.[212] Machine learning research hasn't developed agreed-upon standards for how to quantify properties like catastrophic risk.

Determining which models merit registration under a capabilities-based test is also complicated by the fact that model capabilities can be advanced through post-deployment finetuning,[213] prompting,[214] or integration with additional software tools.[215] The performance improvements from these steps may be substantial.[216] This complicates enforcement: regulators may initially determine a model does not meet capability thresholds, only to discover later that augmentation with specific APIs or the use of a specific prompting technique enables the model to meet the thresholds.

---

[207] There is some disagreement, however, on whether such databases should be public. *Compare* Hadfield et al., *supra* note 199, *with* Blumenthal & Hawley, *supra* note 81.
[208] Hadfield et al., *supra* note 199.
[209] Hadfield et al., *supra* note 199.
[210] Blumenthal & Hawley, *supra* note 81.
[211] Anderljung et al., *supra* note 199.
[212] Toby Shevlane et al., *Model Evaluation for Extreme Risks*, ArXiv (May 25, 2023) https://arxiv.org/pdf/2305.15324.pdf.
[213] *Transfer Learning and Fine-Tuning*, TENSORFLOW, https://www.tensorflow.org/tutorials/images/transfer_learning (last visited Sep. 14, 2023).
[214] Jason Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, ARXIV (Jan. 28, 2022), https://arxiv.org/abs/2201.11903.
[215] *See, e.g.*, *ChatGPT Plugins*, OPENAI (Mar. 23, 2023), https://openai.com/blog/chatgpt-plugins; Anderljung et al., *supra* note 199.
[216] Wei et al., *supra* note 214.

A registration requirement may also encounter feasibility challenges if the eligibility criteria used by regulators poorly captures the intended targets of the system. For instance, criteria may be too broad and therefore require registration of models which do not actually exhibit properties that led regulators to initially impose registration. Alternatively, it may be too narrow, and fail to capture important categories of systems which regulators intended to cover.

For instance, registration proposals which require models of a certain size (measured in parameters) or trained with a certain number of FLOPS (an approximate measure of the computational extensiveness of pretraining) provide an example of this concern.[217] These proposals presume that parameter count and FLOPS are loose proxies for model capabilities, allowing regulators to single out more advanced models in a more standardized way. However, recent research suggests that capabilities exhibited by frontier models can be elicited in smaller models through improved algorithmic choices.[218] Registration systems which use model or training data size as a proxy for capabilities are thus in jeopardy of quickly becoming outpaced by AI progress.[219]

Finally, concerns about overly broad eligibility criteria can arise even for registration systems which target certain domains of use (e.g., healthcare or criminal justice). Registration schemes for these settings must distinguish AI from existing software systems or algorithmic tools. Already, many have observed the inherent difficulty in even defining what "AI" is,[220] and the propensity for certain definitions to inadvertently include benign systems.[221] Thus, registration regimes based on domain use may require regulators to divine boundaries between "new" forms of AI and older data-based tools—for example, the blurry line between AI and clinical decision support systems [222]

[217] Hadfield et al., *supra* note 199.
[218] Rohan Taori et al., *Alpaca: A Strong, Replicable Instruction-Following Model*, STANFORD CRFM (2023), https://crfm.stanford.edu/2023/03/13/alpaca.html.
[219] Dylan Patel et al., *Google: "We Have No Moat, and Neither Does OpenAI"*, SEMIANALYSIS (May 4, 2023), https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.
[220] *See, e.g.,* Lawrence, Cui, & Ho, *supra* note 105, for discussion of how an ambiguous definition of AI may be a contributing factor to the inconsistent publication, as required by EO 13960, of agency AI use case inventories.
[221] Matt O'Shaughnessy, *One of the Biggest Problems in Regulating AI Is Agreeing on a Definition*, CARNEGIE ENDOWMENT FOR PEACE (Oct. 6, 2022), https://carnegieendowment.org/2022/10/06/one-of-biggest-problems-in-regulating-ai-is-agreeing-on-definition-pub-88100.
[222] Reed T. Sutton et al., An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success, NPJ Digital Medicine, February 2020, https://www.nature.com/articles/s41746-020-0221-y.

B. Institutional Feasibility: Registration Regimes Would Face Significant Concerns about Volume, Evasion, and Inter-Agency Coordination

The first question raised by registration is a simple one: do regulators have the capacity to implement and maintain a registration system? The apparent simplicity of registration obscures the resources it necessitates. For instance, effective registration requires regulators to ensure that submitted statements are accurate. In the context of AI, developers may make claims that can only be checked through in-depth audits, and regulators may not have the authority or capacity to perform inspections to verify. And though omissions and deception in registration statements may create liability for developers, to deter such misrepresentations requires regulators to expend the resources investigating and policing non-compliance.

Consider clinical trial registration, which requires clinical trial sponsors to record trial results in a federal database.[223] A study of a sample of trials revealed that almost 55% were in violation of federal reporting requirements even though delayed reporting could accrue thousands of dollars in daily penalties.[224] And despite calls for increased enforcement, FDA and National Institutes of Health (NIH) officials have shied away from any punitive actions.[225]

Registration feasibility also depends on the breadth of the system. If eligibility criteria are too inclusive, regulators risk being overwhelmed with statements. This has two repercussions. First, regulators are less likely to catch errors in individual statements. Second, regulators are more likely to miss registrations corresponding to particularly salient risks. These concerns are significant for AI, given the number of models that may require registration. Huggingface—a repository for the open-source community to share and distribute AI artifacts—has over 120 thousand models.[226] The fact that even a small proportion of these models may require registration—not to mention the population of models not contained on the platform—could overwhelm a registration system.

The dynamic and fast-paced nature of AI development also raises important concerns as to whether the regulators would be able to maintain pace with released AI systems. For instance, regulators who implement a registration system that uses model size thresholds

[223] Charles Piller, FDA and NIH Let Clinical Trial Sponsors Keep Results Secret and Break the Law, Science (Jan. 12, 2020), https://www.science.org/content/article/fda-and-nih-let-clinical-trial-sponsors-keep-results-secret-and-break-law.
[224] *Id*.
[225] *Id*.
[226] *Huggingface Hub Documentation*, HUGGINGFACE (last visited Oct. 1, 2023), https://huggingface.co/docs/hub/index.

as proxies for capabilities may find themselves rapidly revising the threshold downwards, as small models continue to improve.[227]

Relatedly, regulators must also account for attempts to evade registration. Targets often actively attempt to bypass registration, by exploring alternate ways of designing or marketing products.[228] A notable example of this is Nvidia's response to U.S. export controls—to avoid the export ban on powerful microchips, Nvidia simply designed chips with slower processing speeds that fall below the performance threshold, resulting in a game of threshold cat and mouse.[229] Regulators will need to determine when such behavior actually fulfills regulatory goals—because targets are avoiding risky behavior or systems—or undermine them.

Challenges with evasion may also arise if regulators use benchmarks to evaluate capabilities to devise registration requirements. The tendency to train frontier models broadly on all web data has raised concerns that high performance on benchmarks may not be representative of actual performance.[230] Benchmark thresholds may thus capture models which do not actually possess significant capabilities, but merely "cheated" at the evaluation.[231] Regulators may also face the opposite challenge. Developers seeking to avoid registration could subtly modify models to fail benchmark evaluations, while maintaining capabilities.[232]

Questions of *how* to manage and enforce a registration system inevitably give way to *who* should do so. A registration scheme would need to account for existing AI-related regulatory authorities.[233] Thus, legislators would have to determine whether registries

---

[227] Taori et al., *supra* note 218.
[228] *See* JEAN-JACQUES LAFFONT & JEAN TIROLE, A THEORY OF INCENTIVES IN PROCUREMENT AND REGULATION (1993). E.g., Sarah Lykken, We Really Need to Talk: Adapting FDA Processes to Rapid Change, 68 Food & Drug L.J. 357, 374 (201"). ("Any system in which (a) levels of regulation depend on product or transaction characterizations and (b) regulated entities have the capacity for rapid innovation, leaves itself vulnerable to entities characterizing their products or transactions in a way that minimizes regulatory costs, whether or not such characterizations accord with regulatory intent.")
[229] Rita Liao, *Nvidia touts a slower chip for China to avoid US ban*, TECHCRUNCH (Nov. 7, 2022, 8:02 PM), https://techcrunch.com/2022/11/07/nvidia-us-china-ban-alternative/; Ana Swanson, U.S. Tightens China's Access to Advanced Chips for Artificial Intelligence, N.Y. TIMES (Oct. 17, 2023), https://www.nytimes.com/2023/10/17/business/economy/ai-chips-china-restrictions.html.
[230] Arvind Narayanan & Sayash Kapoor, *GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question*, AI SNAKE OIL (Mar. 20, 2023), https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.
[231] Amandalynne Paullada et al, *Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research*, ARXIV (Dec. 9, 2020), https://arxiv.org/abs/2012.05345.
[232] *See generally* Michael F. Stumborg et al., *Goodhart's Law*, Ctr. Naval Analyses (Sep. 1, 2022), https://www.cna.org/reports/2022/09/goodharts-law ("When a measure becomes a target, it ceases to be a good measure.").
[233] *See supra* notes 89–96.

should be managed by agencies already regulating AI or by a single entity across all domains. If registries and adverse event databases are spread across multiple agencies, it may be more difficult to identify macro trends or improve public understanding about AI uses and associated risks without interagency information sharing and coordination processes, which can be challenging to create.[234] However, centralization within one entity may be difficult—no single agency currently has jurisdiction over all of [235] Consistent implementation and maintenance of a registry would require legislators to determine the types of expertise that matter, and to assign responsibility for enforcement. The ability to secure international cooperation in enforcement would also affect the efficacy of a domestic registration regime. Registration proposals should consider which domestic entity would be best placed to emphasize international cooperation.[236]

C. Registration's Tensions: Registration May Reduce Information Asymmetries But Also Undermine Independent Evaluation

As a stand-alone intervention, registration is best positioned to alleviate informational gaps in regulators' understanding of an industry or domain.[237] But because registration systems can be costly to implement and enforce, regulators should clarify whether existing information deficits forestall beneficial regulation. Just as disclosure's benefit is most concrete when it can be associated with private decision-making, registration's benefit is most clear when it can be linked to governmental decision-making.

When combined with other interventions, registration's benefits can manifest in broader ways. First, registration can provide critical infrastructure for other regulatory action, like adverse event reporting systems for foundation models (disclosures).[238] Adverse event reporting frameworks require that regulators be able to identify when different reports refer

---

[234] *See, e.g.,* Administrative Conference Recommendation 2012-5: Improving Coordination of Related Agency Responsibilities, ADMIN. CONF. U.S. (adopted June 15, 2012), https://www.acus.gov/recommendation/improving-coordination-related-agency-responsibilities (explaining that overlapping delegations and a "shared regulatory space" can create "may produce redundancy, inefficiency, and gaps," and "underappreciated coordination challenges"); *Leading Practices to Enhance Interagency Collaboration and Address Crosscutting Challenges*, U.S. GOV'T ACCOUNTABILITY OFF. (May 2023), https://www.gao.gov/assets/gao-23-105520.pdf.

[235] *See supra* pp. 22–23.

[236] For a longer discussion of the international cooperation challenges, *see infra* Section IV.B.

[237] Grant Wilson, *Minimizing Catastrophic and Existential Risks From Emerging Technologies Through International Law*, 31 VA. ENV'T. L.J. 307, 318 (2013).

[238] The Biden administration, along with major AI developers, has already sought to promote public assessments of AI systems through red-teaming. Press Release, White House, FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety (May 4, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

to the same underlying entity.[239] Registration regimes provide a mechanism for naming and identifying which systems meet certain criteria. Without registration, regulators may struggle to track which systems reports refer to, which developers to follow up with, and whether event reports refer to a current system. In the parlance of computer scientists: registration provides regulators with a *schema* for organizing and collecting different types of information.

Second, registration can enhance the effectiveness of other regulatory interventions. For instance, there is empirical evidence to suggest that the introduction of registration requirements lowered misreporting by hedge funds.[240] Mandatory registration for AI systems may similarly lead developers to create better internal compliance structures or evaluation practices. Borrowing a trick from clinical trial registration, regulators might also consider whether registration can be used to systematize how developers perform publicly reported evaluations. For instance, regulators could require developers to register pre-deployment tests and report basic information regarding the test datasets used, metrics, and other evaluation protocols. Such practices could assuage concerns regarding evasion or manipulation.[241]

In practice, registration systems can appear to combine elements of mandatory disclosure and licensing. Though information produced through registration need not be made public, registration regimes that release information to the public also achieve the goals of disclosure.[242] And though registration need not involve government review of the statement, many regimes require government approval, thereby empowering the government to take the role of a licensor.

Third, though registration is intended to enhance the quality and volume of information available to regulators, it can have precisely the opposite effect (vertical misalignment) if poorly implemented. Today, a significant fraction of our understanding of large foundation models comes from research and open/open-source efforts to develop and understand these models.[243] Registration requirements that burden researchers or open-source communities

[239] Adverse Event Detection, Processing, and Reporting, https://www.ncbi.nlm.nih.gov/books/NBK208615/.
[240] Colleen Honigsberg, Hedge Fund Regulation and Fund Governance: Evidence on the Effects of Mandatory Disclosure Rules, 57 Journal of Accounting Research 845 (2016).
[241] Tom Simonite, *Why and How Baidu Cheated an Artificial Intelligence Test*, WIRED (June 4, 2015), https://www.technologyreview.com/2015/06/04/72951/why-and-how-baidu-cheated-an-artificial-intelligence-test/.
[242] For instance, the FDA requires that clinical trials be registered, and shares this data publicly on clinicaltrials.gov, for the express purpose of educating patients and doctors about clinical research.
[243] Will Knight, The Myth of "Open Source" AI, Wired (Aug. 24, 2023), https://www.wired.com/story/the-myth-of-open-source-ai/.

could functionally operate to slow public understanding of these systems[244] In short: if it is information that regulators desire, then researchers should be mobilized, not encumbered.

Fourth, registration may also be in tension with other regulatory goals (horizontal misalignment). Sector-based registration schemes are often premised on the idea that sensitive areas (e.g., benefits distribution or healthcare) should be slow to deploy AI. Yet, the incorporation of AI into these sectors may be essential for maintaining national competitiveness.[245] AI here could enhance government efficiency, allowing agencies to better manage economic schemes vital to national health and welfare.[246]

Assessing registration's suitability to address specific harms also necessitates the identification of existing regulatory gaps or baseline risks. For instance, although LLM registration is cited as necessary for the nonproliferation of bioweapons, existing resources can also furnish relevant information—including Google search or public libraries neither of which require registration to use. Given the broad accessibility of these resources, regulators might reasonably conclude that a more fruitful focus is restricting access to materials essential to developing bioweapons, irrespective of the point of access.

Finally, the tendency to bundle registration with more punitive tools means that registration may functionally alter both who gets to build, and benefit, from AI. Onerous registration requirements for developing AI systems will concentrate development amongst large organizations and reduce the ability for smaller or emerging companies to compete. Similarly, registration requirements which steer AI development away from sensitive settings like healthcare may only deprive those who may benefit the most of AI's rewards.

V. Licensing

[244] Nitasha Tiku et al, *Google Shared AI Knowledge with the World — Until ChatGPT Caught Up*, WASHINGTON POST (May 5, 2023), https://www.washingtonpost.com/technology/2023/05/04/google-ai-stop-sharing-research/.
[245] Commission on Artificial Intelligence Competitiveness, Inclusion, and Innovation, US Chamber of Commerce (2023), https://www.uschamber.com/assets/documents/CTEC_AICommission2023_Report_v5.pdf.
[246] Daniel E. Ho, *Opportunities and Risks of Artificial Intelligence in the Public Sector* (May 16, 2023), https://www.hsgac.senate.gov/wp-content/uploads/Testimony-Ho-2023-05-16-1.pdf.

Licensing regimes[247] authorize entities or individuals to conduct or engage in an activity that is otherwise legally prohibited.[248] Thus, regulatory licensing goes beyond registration by creating a system of more direct regulatory gatekeeping through a combination of standards and evaluations paired with the threat of sanctions for violations.[249] Common goals of licensing regimes are to increase public health and safety, ensure a minimum quality of professional competency, and prevent fraud, abuse, and evasion of national security-related policies (e.g., export controls, sanctions).[250]

Although licensing and registration regimes are sometimes discussed interchangeably, we distinguish them by focusing on the burden placed on regulated entities and the primary motivation of regulators. Licensing typically requires a government entity to engage in significant oversight, review, and deliberation prior to granting a license, and is often employed where regulators seek to maintain minimum quality standards (e.g., professional licensing) or address scarcity, either naturally (e.g., limited natural resources) or to limit an activity (e.g., to minimize pollution).[251] In comparison, registration operates more like a check-the-box activity intended to ensure government can monitor a certain limited subset of activities and respond to adverse events. Of course, this distinction is often blurry, with

[247] We are not examining contractual licensing agreements between parties about the use of intellectual property.
[248] *E.g.*, *OFAC License Application Page*, U.S. Dep't of the Treas., https://ofac.treasury.gov/ofac-license-application-page ("A license is an authorization from OFAC to engage in a transaction that otherwise would be prohibited."); *U.S. Export Licenses Navigating Issues & Resources*, Int'l Trade Admin., https://www.trade.gov/us-export-licenses-navigating-issues-and-resources ("An export license is a government document that authorizes or grants permission to conduct a specific export transaction (including the export of technology)."); *Licensing*, U.S. Nuclear Regul. Comm'n, https://www.nrc.gov/about-nrc/regulatory/licensing.html (explaining that a license "authorizes an applicant" to construct and operate commercial reactors and fuel cycles, possess and use nuclear materials and waste, and construct and operate waste disposal sites, among other activities); Ryan Nunn, *How occupational licensing matters for wages and careers*, BROOKINGS (Mar. 15, 2018), (explaining that occupational licensing is "the legal requirement that a credential be obtained in order to practice a profession"); *Types of Licenses*, FDA, https://www.fda.gov/science-research/licensing-and-collaboration-opportunities/types-licenses (detailing the types of licenses the FDA offers commercial partners to develop and market FDA-created technologies).
[249] *E.g.*, *Public Involvement in Licensing*, U.S. Nuclear Regul. Comm'n, https://www.nrc.gov/about-nrc/regulatory/licensing/pub-involve.html; *Licensing*, *supra* note 248; *OFAC Licenses*, U.S. Dep't of the Treas., https://ofac.treasury.gov/faqs/topic/1506; MARC LABONTE, WHO REGULATES WHOM? AN OVERVIEW OF THE U.S. FINANCIAL REGULATORY FRAMEWORK, Cong. Rsch. Serv., R44918 (Mar. 10, 2020), https://sgp.fas.org/crs/misc/R44918.pdf.
[250] *E.g.*, *Occupational Licensing: A Framework for Policymakers*, WHITE HOUSE 2 (July 2016), https://obamawhitehouse.archives.gov/sites/default/files/docs/licensing_report_final_nonembargo.pdf ("When designed and implemented appropriately, licensing can benefit practitioners and consumers through improving quality and protecting public health and safety.")
[251] Breyer, *supra* note 57, at 71.

some nominal registration requirements functioning as licensing regimes[252]–highlighting our point that distinctions between regulatory regimes can functionally collapse.[253]

Licensing regimes vary in the degree of prescriptive requirements. Occupational licensing, for example, often requires individuals to meet certain education, training, and testing requirements.[254] Although occupational certification authorizes individuals to practice a particular line of work after achieving a certain educational or skill level, occupational licensing is more rigorous and requires an applicant to apply for a license, provide additional information, pay a fee, and in some professions (e.g., law) pass character, fitness, ongoing education, or other background checks.[255] Gun licenses required in some states similarly require individuals take and pass certified firearm safety courses and meet certain background requirements (e.g., no felony convictions), pay a fee, and get interviewed by a law enforcement or government official.[256] A license to sell a vaccine includes several "pre-market" requirements: an entity must submit to the FDA extensive information about the vaccine, the manufacturer, preclinical and clinical studies, and draft vaccine labeling and await extensive FDA review of the information provided and, in some cases, inspection of the manufacturer.[257] Thus certain licensing regimes clearly delineate, and even actively control the substance of, requirements.

Other licensing regimes that govern exporting or other sensitive activities provide the government agency more leniency to grant licenses "after a careful review of the facts"[258] or on a "case-by-case basis."[259] Although rare, some licensing regimes (e.g., Nuclear Regulatory Commission licensing of commercial reactors) provide an opportunity for the

[252] *See, e.g.*, *About Pesticide Registration*, EPA, https://www.epa.gov/pesticide-registration/about-pesticide-registration (last visited Oct. 6, 2023) (registration process which is evaluated for approval based on a detailed cost-benefit analysis of the pesticide's use).
[253] *See infra* discussion in Section VII.A.
[254] *Occupational Licensing: A Framework for Policymakers*, *supra* note 250, at 12.
[255] *Id.* at 44, Nunn, *supra* note 248; Paul J. Larkin, *Public Choice Theory and Occupational Licensing*, 39 HARVARD J.L. & PUB. POL'Y 209, 210 (2016).
[256] *Concealed Pistol Licenses (CPL)*, Clerk/Register of Deeds Washtenaw County-Michigan, https://www.washtenaw.org/521/Concealed-Pistol-Licenses; *Firearms License & Renewals*, The Town of Concord, Massachusetts, https://concordma.gov/308/Firearms-License-Renewals.
[257] The Biologics License Application (BLA) Process Explained, The FDA Group (Mar. 28, 2022), https://www.thefdagroup.com/blog/2014/07/test-the-biologics-license-application-bla-process/; Biologics License Applications (BLA) Process (CBER), FDA, https://www.fda.gov/vaccines-blood-biologics/development-approval-process-cber/biologics-license-applications-bla-process-cber (explaining that a Biologics License Application is a "a request for permission to introduce, or deliver for introduction, a biologic product into interstate commerce").
[258] *U.S. Export Licenses Navigating Issues & Resources*, *supra* note 248 ("Export licenses are issued by the appropriate licensing agency after a careful review of the facts surrounding the given export transaction.").
[259] *OFAC License Application Page*, *supra* note 248 ("OFAC will consider the issuance of specific licenses on a case-by-case basis when a general license provision is not available.")

public to participate in agency decision-making by submitting comments or participating in agency hearings.[260]

Proposals for AI licensing regimes are intended to ensure responsible and skilled development and use of AI products, either by licensing companies or practitioners, or through approval of the development or deployment of systems themselves. The belief that unhindered AI development and deployment creates risks to public safety and consumer protection commonly animates calls for AI licensing. Proposals vary in terms of what type of activity is being regulated (development or deployment of AI), what entity is subject to the regulation (organizations or individuals), and what type of AI model must be licensed.

Most proposals would require an organization or individual to obtain a license before *deploying* an AI model that poses a certain degree of risk to consumers or society. In addition to requiring companies to register the *development* of "sophisticated general-purpose AI models" and AI used in "high-risk situations,"[261] the Blumenthal-Hawley "Bipartisan Framework for U.S. AI Act" would establish an "independent oversight body" to grant licenses to companies that seek to *deploy* such models. To obtain a license, companies would have to provide certain information about the models (i.e., register the models), maintain certain compliance programs (risk management, data governance, pre-deployment testing, and adverse incident reporting), and be subject to audits by the oversight body.[262]

The private sector has also advocated for licensing. OpenAI CEO Sam Altman proposed licensing as part of a comprehensive regulatory framework in his Senate testimony.[263] Analogizing to regulation of pharmaceutical drugs, OpenAI researchers have suggested that, if AI models "pose risks to public safety above a high threshold of severity," frontier AI developers should obtain a "license to widely deploy" the frontier AI model upon demonstrating compliance with safety standards.[264] One law review article proposed "An FDA for Algorithms," including the creation of an agency that could conduct pre-market reviews, such as safety studies, and approve algorithms before deployment.[265] Licensing

[260] *Public Involvement in Licensing*, *supra* note 274; *Licensing*, *supra* note 248.
[261] *Supra* Section IV.
[262] Blumenthal & Hawley, *supra* note 81.
[263] *Oversight of AI: Rules for Artificial Intelligence: Hearing Before the Subcomm. On Privacy, Technology, and the Law of the Senate Judiciary Comm.*, 118th Cong. (2023) (transcript available at https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/) (statement of Samuel Altman, CEO, OpenAI) [hereinafter *Hearing on Rules for AI*].
[264] Anderljung et al., *supra* note 199, at 20.
[265] Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 111 (2017).

requirements for testing prior to release and distribution would mirror existing FDA approval processes for medical devices (including AI-enabled devices[266]).

Proposals for licensing the *development* of AI models are motivated by concerns that high-risk AI may be stolen or leaked, become available through small-scale deployment intended to test the AI models, or may evade regulation, particularly where the models are never intended for wide deployment.[267] Thus, OpenAI researchers argue that licensing the *development* of frontier models may be necessary and could be contingent upon developers having security and theft-protection measures, conducting risk assessments before training runs, and adopting risk management practices like incident registers.[268]

Other AI licensing proposals draw more similarities to occupational licensing, focusing on the *ability of the entity or individual* to develop and deploy AI, instead of the development or deployment of a particular AI model. For example, Senators Elizabeth Warren and Lindsey Graham's Digital Consumer Protection Commission Act proposes the creation of an Office of Licensing for Dominant Platforms within an independent regulatory Digital Consumer Protection Commission that would require generative AI platform[269] companies that are "dominant"—defined as meeting a minimum monthly active users and net annual sales threshold—to obtain a license to operate.[270] C-suite executives would be required to annually certify their compliance with numerous mandates, including disclosure requirements, prohibitions on anti-competitive practices and foreign access to data, privacy protections, and commitments to uphold duties of care and mitigate risks (e.g., discrimination, addictive behaviors).[271]

---

[266] Wu et al., *supra* note 89.
[267] Anderljung et al., *supra* note 199, at 20–21 (explaining the rationale for development licensing including that certain models "may be used to, for example, develop intellectual property that the developer then distributes via other means").
[268] *Id.*
[269] Part C defines a platform to mean "a website, online or mobile application, operating system, online advertising exchange, digital assistant, or other digital service that . . . (C) enables user searches or queries that access or display a large volume of information." Sec. 2002, 3), www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf.
[269] Microsoft's blueprint for AI regulation calls for licensing of both large models and the data centers in which they are hosted. MICROSOFT, GOVERNING AI: A BLUEPRINT FOR THE FUTURE 20 (2023), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.
[269] Digital Consumer Protection Commission Act of 2023, S. 2597, 118th Cong.
[270] *Id.*; Lenhart, *supra* note 93.
[271] Digital Consumer Protection Commission Act of 2023, *supra* note 269, § 2604; Title-by-Title Summary of the Digital Consumer Protection Commission Act, Office of Senator Elizabeth Warren, https://www.warren.senate.gov/imo/media/doc/DCPC%20Section-By-Section.pdf; Press Release, Senator Elizabeth Warren, *Warren, Graham Unveil Bipartisan Bill to Rein in Big Tech* (July 27, 2023), https://www.warren.senate.gov/newsroom/press-releases/warren-graham-unveil-bipartisan-bill-to-rein-in-big-tech.

In addition, some organizations have called for developing professional standards or licensure requirements in data science and machine learning to address safety concerns, strengthen accountability, and promote ethical conduct.[272]

A. Technical Feasibility: Defining Standards Agnostic to Application is Challenging

AI licensing requirements suffer from the same technical challenges as registration regimes, but also face additional challenges arising from the need to develop, often *ex ante*, criteria for granting and revoking licenses. Challenges that regulators face in determining which systems require registration are only exacerbated in the licensing context as most proposals envision a smaller class of AI models subject to the more burdensome requirements.[273] Thus, questions about determining which capabilities actually pose risk, and determining how to measure or proxy those capabilities would become even more challenging for regulators to navigate.

A second challenge is that pre-market approval standards and evaluation criteria—for development and deployment licenses—are exceptionally challenging to define independent of knowledge about the context or application for the AI model. Pre-market standards are most effective when they can be tailored to capture a technology's performance as it is used.[274] For instance, crash tests are designed to mimic accident trajectories common to real-world crashes.[275] However, for many classes of machine learning models—most notably "foundation models" like GPT-4 or CLIP—the full spectrum of use cases may not be known at the time of creation. This is because these types of AI systems often enable numerous applications. Unlike conventional AI models—which are developed to perform one task—foundation models are trained to learn common

[272] *E.g.*, Danish Contractor et al., *Behavioral Use Licensing for Responsible AI*, PROC. 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 778 (2022); Martin Kandlhofer & Gerald Steinbauer, *A Driving License for Intelligent Systems*, 32 PROC. AAAI CONF. ON A.I. 7954 (2018); Kathirvel Kumararaja, Do We Need Licensing for Working on Artificial Intelligence Technology? (May 20, 2023), https://www.linkedin.com/pulse/do-we-need-licensing-working-artificial-intelligence-kathirvel/;

[273] For the sake of brevity, we refer readers to the discussion of these challenges in Section IV.A.

[274] See, e.g., Keith Barry et al, *The Crash Test Bias: How Male-Focused Testing Puts Female Drivers at Risk*, CONSUMER REPORTS (Oct. 23, 2019), https://readwise.io/reader/document_raw_content/89573668 (observing how a focus on crash testing dummies which capture male anatomy could explain differences in safety for men and women in real world car crashes).

[275] *See, e.g.*, *Biomechanics*, Nat'l Highway Traffic Safety Admin., https://www.nhtsa.gov/research/biomechanics (explaining that NHTSA conducts "cooperative and collaborative research with other organizations" including "collection and analysis of rea-world injury data, development and evaluation of advanced testing and simulation tools such as crash test dummies" to improve motor vehicle safety); *Ratings*, Nat'l Highway Traffic Safety Admin., https://www.nhtsa.gov/ratings; *Crashworthiness*, Nat'l Highway Traffic Safety Admin., https://www.nhtsa.gov/research-data/crashworthiness.

patterns in different modalities of data (e.g., text or images). They are, as many researchers have noted, inherently "taskless,"[276] raising the question: how do regulators define standards for a technology not engineered towards a specific application?[277]

A third challenge is that deployed AI systems are often subject to frequent updates. Although this challenge is also present in disclosure, registration, and auditing regimes, dealing with updates is particularly important in the context of licensing given its gatekeeping function. AI model updates serve important purposes, allowing developers to address drifts in data distribution, changing real-world conditions, and identified bugs. Updates to systems are nontrivial and can meaningfully change model behavior, models themselves, or risks posed by models (e.g., potential for misuse or vulnerability to attacks). Regulators must thus define re-licensing criteria–when is an update so substantial as to require developers to "reapply" for a license?

B. Institutional Feasibility: Challenges with Supervision and Enforcement

Compliance with licensing requirements will hinge on a variety of institutional factors that center around the capacity for government to approve and revoke licenses.

First, establishing and implementing a licensing regime requires expertise and capacity to define criteria, approve licenses, monitor for noncompliance, and revoke licenses, as necessary. Professional licensing regimes also often require identifying and delineating skills and knowledge requirements and certifying courses or examinations. Current licensing proposals seem to coalesce around the creation of *one* entity to oversee an AI licensing regime, but licensing targeted at high-risk uses must account for the fact that these sectors are already subject to regulatory oversight, by agencies which have acquired their own significant expertise.[278]

Particularly instructive about this institutional feasibility challenge is the fact that we already have a licensing scheme in place: the FDA's pre-existing medical device licensing

[276] *See, e.g.*, Christina Montgomery et al., *A Policymaker's Guide to Foundation Models*, IBM (May 1, 2023), https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models; Bommasani et al., *supra* note 39.
[277] It's also helpful to note that because our understanding of these systems is still in its infancy, researchers are still learning about potential applications. In fact, a meaningful portion of AI research today is devoted to the question of understanding: where do foundation models work well, and where do they fail? *E.g.*, Neel Guha et al., *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*, ARXIV (Aug. 20, 2023), https://arxiv.org/abs/2308.11462.
[278] *See supra* notes 89–96.

regime has approved, as of October 2022, over 500 AI medical devices.[279] And although the vast majority were approved in the last five years, the first AI/ML-enabled medical device was approved in 1995.[280] By regulating the use of AI in medical devices and implementing pre-market approval programs for drugs and biologics like vaccines, the FDA has expertise and experience relevant for establishing a licensing regime.

However, the existing medical device approval processes will need to adapt to allow for the more frequent and population-specific updates that may be necessary for AI-enabled systems than, say, electronic medical record systems or MRI machines.[281] Doing so well will require a careful consideration of policy design as regulators will face a trade-off between the harms that may be prevented by exercising more control over system approvals and those that might be introduced by creating barriers to model adaptations across settings. An AI system trained in one setting may see appreciable performance degradations over time or when applied to new populations. Approval processes that treat AI systems as static and universally applicable may perform poorly in novel settings.[282]

For medical devices, FDA has begun to explore flexible regulation through "Predetermined Change Control Plans" that would limited updates over time to be covered by an initial approval, but not, for instance, relative to setting or patient population.[283] How best to navigate this trade-off is unclear, and likely highly application-dependent, but is a critical dimension for policymakers to explore: To what extent should regulators provide flexibility for device updates, and across what dimensions (e.g., over time, setting, modeling target, model structure)? What sort of guardrails need to be put in place to ensure updated models meet some baseline performance and fairness criteria, and how should those criteria be set?

---

[279] Dave Fornell, FDA has now cleared more than 500 healthcare AI algorithms, HEALTHEXEC (Feb. 6, 2023), https://healthexec.com/topics/artificial-intelligence/fda-has-now-cleared-more-500-healthcare-ai-algorithms; FDA, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (current as of Oct. 5, 2022), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. The exact number of devices approved as of October 2, 2023, is likely to be much more.

[280] As of October 2, 2023, 91 were approved in 2022, 115 in 2021, 102 in 2020, 77 in 2019, 63 in 2018. FDA, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (current as of Oct. 5, 2022), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[281] Wu et al., *supra* note 89.

[282] Bommasani et al., *supra* note 39, at 109–113 (discussing the challenges in AI related to so-called "distribution shifts").

[283] Ctr. For Devices & Radiological Health, FDA-2022-D-2628, Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions, U.S. Food & Drug Admin (2023), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial.

AI systems in other high-stakes domains could be subjected to similar review. However, the differences between narrowly scoped medical AI systems and frontier models is such that the FDA's premarket approval approach and timelines may not be appropriate for the latter class of systems.[284] For example, the flexibility for updates in the FDA's "Predetermined Change Control Plans," may be dependent on the policy domain (e.g., health devices vs. self-driving cars). The FDA is thus a bellwether for AI licensing, where the same issue of approval of model evaluation, adaptation, and performance will challenge regulatory capacity.

Each agency with AI-related regulatory authority will undoubtedly be implicated in the enforcement of any registration scheme, and perhaps have developed their own perspectives on how best to navigate AI risks and benefits. At a minimum, existing licensing will need to adapt. But a single entity overseeing licensing may struggle to leverage agency-specific expertise without creating needless duplication.[285] And, again, there are significant concerns about the government's ability to hire, train, and retain technical talent, with additional resources undoubtedly necessary for an entity to define licensing criteria, review applications, and grant and revoke licenses.

Second, attempts to limit the volume of AI models subject to the licensing regime (e.g., licensing only the deployment of frontier AI models that pose significant risk) could become meaningless, potentially overwhelming the licensing entity or entities. As discussed above, technical advancements and the democratization of AI knowledge and resources means that sophisticated systems can be run or developed from even basic devices. The architectures for many sophisticated systems have also been published in openly accessible papers. The software libraries necessary for training models—and the data to train them—are freely available. And even if large proprietary models like GPT-4 require more compute than any one person can access, algorithmic innovations allow for developing "small" models which can beat large proprietary models, at minimal cost. Thus, the number of AI systems that could meet licensing criteria, particularly given technological innovations, is not likely to remain constant or decrease. License renewals—particularly if renewals are required at frequent intervals[286]—may also increase the burden facing a licensing entity. Implementation and enforcement of the licensing regime would thus be no small task.

---

[284] For example, the FDA's median review time for standard and priority drug applications was 2.8 years from 1986 to 1992. The median review time for standard drug applications decreased to 10.1 months in 2018, but researchers found the time savings may derive from a reliance on less evidence. Sydney Lupkin, *FDA Approves Drugs Faster Than Ever But Relies On Weaker Evidence, Researchers Find*, NPR (Jan. 14, 2020), https://www.npr.org/sections/health-shots/2020/01/14/796227083/fda-approves-drugs-faster-than-ever-but-relies-on-weaker-evidence-researchers-fi.

[285] *See supra* discussion in Section IV.B.

[286] *See supra* discussion in Section V.A.

Third, an AI licensing regime is particularly susceptible to evasion. Although violations of non-AI licensing for certain activities (e.g., flying airplanes or using pesticides) or professions are well-documented,[287] preventing the unlicensed development or deployment of AI, particularly outside of the United States, will be particularly difficult. Although not a licensing regime, U.S. and global attempts to prevent the exportation of "dual-use" technologies (e.g., facial recognition) used for both military and civilian purposes is a useful comparison. Challenges tracking and preventing the exportation of *software* have led the international community to focus on export controls targeting *hardware* used to power certain technologies. However, companies may alter their products to avoid compliance with export controls.[288] Similar challenges will likely face AI licensing. Like foreign demand for chips, domestic demand for AI models may incentivize overseas entities to evade burdensome or unattainable licensing. Preventing non-U.S. entities and individuals from developing and deploying frontier AI will thus be challenging, even where U.S. allies and partners coordinate or adopt similar licensing regimes.

C. Licensing's Tensions: Anti-Competitive and Incumbent Enhancing?

Licensing regimes are frequently used in contexts where activities that harm the public at large or individual consumers result from information asymmetries—because it is costly or difficult to obtain information about the regulated activity.[289] Consumers may not be able to identify ex ante whether an unlicensed doctor will perform a safe surgery and governments may not trust that any company running a nuclear fuel cycle facility will have sufficient safety protocols, despite the company's claims. Licensing regimes thus promote transparency around approved tools and practitioners and ensure compliance with standards[290] through sanctions or revocation of licensure for misconduct.

---

[287] Breyer, *supra* note 57, at 71.

[288] *See* Liao, *supra* note 229. More recent proposals to enforce export controls by restricting access to U.S. cloud computing services would require such services to implement "Know Your Customer" controls, which could, theoretically, become more difficult as models require less compute to achieve similar capabilities. For more discussion about cloud-based export controls, see HANNA DOHMEN ET AL., CONTROLLING ACCESS TO ADVANCED COMPUTE VIA THE CLOUD: OPTIONS FOR U.S. POLICYMAKERS, PART I (2023), https://cset.georgetown.edu/article/controlling-access-to-advanced-compute-via-the-cloud/; HANNA DOHMEN ET AL., CONTROLLING ACCESS TO COMPUTE VIA THE CLOUD: OPTIONS FOR U.S. POLICYMAKERS, PART II (2023), https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/.

[289] DEPT. OF TREASURY OFF. ECON. POL'Y, COUNCIL OF ECON. ADVISERS & DEP'T OF LABOR, OCCUPATIONAL LICENSING: A FRAMEWORK FOR POLICYMAKERS 7 (2015), https://obamawhitehouse.archives.gov/sites/default/files/docs/licensing_report_final_nonembargo.pdf.

[290] Standards may be around training, performance, and adherence to an ethical code of conduct (e.g., via individual examination, institutional accreditation, or system/device approval process).

Licensing regimes can be conceptually hard to distinguish from other interventions, and in fact may be strengthened by combining with other AI regulatory regimes. Licensing is often used in scenarios where greater transparency to the public (disclosures) or to the government (registration) is deemed an insufficient safeguard against potentially harmful activity. However, pre-conditioning license approval on registration or disclosure is a common approach in non-AI-focused regulation and in AI licensing proposals,[291] as it enables the government to make more informed licensing decisions.

Pre-market approval procedures, particularly where a government entity like the FDA or NRC reviews or inspects testing and research, can also begin to resemble audit requirements. In the AI context, Anthropic committed in its "Responsible Scaling Policy" not to deploy models that exhibit catastrophic misuse potential, analogizing its risk management framework tiered by "AI Safety Levels (ASL)" to automotive or aviation "pre-market testing and safety" practices that "rigorously demonstrate the safety of a product before it is released onto the market".[292] Anthropic notes it is "developing evaluations for [bioweapons] risks with external experts" and suggests that higher ASL levels may warrant "verifiability" of its internal testing and risk management practices "by external audits."[293] Anthropic's suggestion that deployment should be conditioned on pre-market testing with external verification could easily translate into a licensing regime with auditing requirements.

Technical and institutional feasibility challenges to determining standard and pre-market testing could be addressed through approaches already being explored for improving AI trustworthiness. For example, EU regulators have proposed the creation of "regulatory sandboxes" to test new products prior to release.[294] Regulatory sandbox pilots in the Fintech space not only informed regulation and improved regulator-industry

[291] *See supra* discussion in Section V. Export controls around defense articles provides an explicit example of paring licensing and registration: "The Arms Export Control Act requires that all manufacturers, exporters, temporary importers, and brokers of defense articles . . . are required to register with the Directorate of Defense Trade Controls (DDTC). . . . It is primarily a means to provide the U.S. Government with necessary information on who is involved in certain ITAR controlled activities and does not confer any export or temporary import rights or privileges. Registration is generally a precondition for the issuance of any license or other approval and use of certain exemptions." *Who Must Register*, DIRECTORATE OF DEF. TRADE CONTROLS, https://www.pmddtc.state.gov/ddtc_public/ddtc_public?id=ddtc_kb_article_page&sys_id=7110b98edbb8d30044f9ff621f96192d (last visited Oct. 2, 2023).
[292] *Anthropic's Responsible Scaling Policy*, ANTHROPIC (Sep. 19, 2023), https://www.anthropic.com/index/anthropics-responsible-scaling-policy.
[293] ANTHROPIC'S RESPONSIBLE SCALING POLICY, VERSION 1.0 (ANTHROPIC 2023), https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf, at 7, 21.
[294] EUR. PARLIAMENTARY RSCH. SERV., ARTIFICIAL INTELLIGENCE ACT AND REGULATORY SANDBOXES (2022), https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf.

communication, but also spurred innovation and facilitated international harmonization.[295] Although there are many challenges,[296] piloting regulatory sandboxes prior to implementation of a licensing regime would enable both businesses and regulators to learn about new technology, model capabilities, and unanticipated risks before broad deployment.[297]

AI licensing regimes, however, may suffer from significant vertical and horizontal misalignment. First, we have discussed at length how technical and institutional challenges may make it nearly impossible for licensing to reduce risks posed by foundation models. As smaller models—trained on less parameters or with significantly less FLOPs[298]—continue to match performance of larger models, licensing criteria conditioned on compute may fail to include AI models of regulatory interest. Indeed, the licensing target may not even cause the risk in question—as illustrated by our prior discussion about the risk of bioweapons emanating not from generative AI but from poorly-regulated laboratories. The Warren-Graham proposed Digital Consumer Protection Commission Act is even more susceptible to regulatory mismatch in that it only requires licensing for companies with large user bases. OpenAI's relative obscurity before releasing ChatGPT demonstrates that pre-existing market power is not a prerequisite to deploying a transformative AI model. And some evidence in occupational licensing indicates diminishing returns in service quality as licensing requirements become increasingly stringent.[299]

Second, the potential of licensing to undermine competition, raise costs to consumers, enable industry capture, and gatekeep professions indicates AI licensing would create horizontal misalignment. Literature examining licensing and significant pre-market approval processes outside of the AI context indicates that licensing can create barriers to entry by increasing the cost of production or labor. For example, the pharmaceutical

[295] REGULATORY SANDBOXES IN ARTIFICIAL INTELLIGENCE, NO. 356, OECD, 16-17(2023), https://doi.org/10.1787/8f80a0e6-en.
[296] *Id.* at 17-18.
[297] *Id.*; Carlos Muñoz Ferrandis et al., *Regulatory Sandboxes Can Facilitate Experimentation in Artificial Intelligence*, OECD.AI (May 31, 2023), https://oecd.ai/en/wonk/sandboxes.
[298] *See* e.g., ThirdAI claims to have performed as well as GPT2-XL despite being pre-trained only on CPUs and with 160 times more efficiency (as measured by FLOPs), although it had 1 billion more parameters and trained for 10 more days. *Introducing the World's First Generative LLM Pre-Trained Only on CPUs: Meet ThirdAI's BOLT2.5B*, THIRDAI BLOG (Sept. 18, 2023), https://medium.com/thirdai-blog/introducing-the-worlds-first-generative-llm-pre-trained-only-on-cpus-meet-thirdai-s-bolt2-5b-10c0600e1af4.
[299] *See e.g.,* Morris M. Kleiner et al*., Relaxing Occupational Licensing Requirements: Analyzing Wages and Prices for a Medical Service*, NAT'L BUREAU ECON. RSCH. (2014), https://www.nber.org/papers/w19906; John Manuel Barrios, Occupational Licensing and Accountant Quality: Evidence from the 150-Hour Rule, Becker Friedman Institute for Research in Economics Working Paper No. 2018-32 (Mar. 23, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2893909.

industry is notorious for high barriers to entry and limited competition.[300] One study found that the median cost to bring new therapeutic drugs and biological agents to market was $985 million.[301] And another found that novel therapeutic devices only faced meaningful competition from incumbents.[302] Concerns about the market power of current large, tech companies and AI companies that benefit from first-mover's advantage are well-known. AI licensing that places significant pre- and post-market burdens on companies may be prohibitively costly for smaller developers.[303]

Licensing the development or deployment of AI thus has the potential to concentrate economic power in the hands of a few large companies, restricting access to cutting-edge technology and potentially undermining the goals of both promoting representation in the field and maintaining a competitive edge in the global market. Licensing may heighten market concentration by advantaging more established incumbents who can more easily bear the licensing costs.[304] Concentration of market power could even exacerbate other harms arising from AI or undermine human values and regulatory objectives these policies aim to promote. If licensing requirements are too onerous, the regulations could function as a (partial) ban, in practice. Stifling innovation will certainly exacerbate concerns about geopolitical competition, particularly if other nations do not similarly limit their innovation. Licensing also creates tradeoffs between openness and control. Open access may provide for less control by enabling individuals with bad intentions or insufficient training to more easily access resources, but it may also increase the likelihood that critical issues with the technology are identified after release. Licensing may make it harder for users to expose harms, especially considering how openness provided mechanisms for discovering and addressing cybersecurity risks.[305] And these potential negative

[300] Patricia M. Danzon, *Competition and Antitrust Issues in the Pharmaceutical Industry* (2014), https://faculty.wharton.upenn.edu/wp-content/uploads/2017/06/Competition-and-Antitrust-Issues-in-the-Pharmaceutical-IndustryFinal7.2.14.pdf.
[301] This study looked at 63 of 355 new products approved by the US Food and Drug Administration between 2009 and 2018. Olivia J. Wouters et al., *Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018*, 323 JAMA 9 (2020), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7054832/.
[302] Vinay K. Rathi et al., *Market Competition Among Manufacturers of Novel High-Risk Therapeutic Devices Receiving FDA Premarket Approval Between 2001 and 2018*, 5 BMJ SURG. INTERV. HEALTH TECH. 1 (2022), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9923248/.
[303] *See e.g.*, Corynne McSherry, Generative AI Policy Must Be Precise, Careful, and Practical: How to Cut Through the Hype and Spot Potential Risks in New Legislation, EFF (July 7, 2023), https://www.eff.org/deeplinks/2023/07/generative-ai-policy-must-be-precise-careful-and-practical-how-cut-through-hype; Sarah Myers West & Jai Vipra, Computational Power and AI: Comment Submission, AI NOW INSTITUTE (June 22, 2023), https://ainowinstitute.org/publication/policy/computational-power-and-ai.
[304] *Digital Health Software Precertification (Pre-Cert) Pilot Program*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-pilot-program (last updated Sep. 26, 2022).
[305] Jeremy Howard, *AI Safety and the Age of Dislightenment*, FAST.AI (July 10, 2023), https://www.fast.ai/posts/2023-11-07-dislightenment.html.

externalities are likely to be more substantial with licensing regimes relative to disclosure and registration policies.

The benefits of occupational licensing in particular are debated. Evidence about licensing's impact on the regulated professions is mixed—with studies claiming licensing has both a positive and negative effect on wages and employment.[306] A 2015 White House report found that occupational licensing reduces employment in licensed occupations and reduces the wages of unlicensed workers relative to licensed workers with similar levels of experience and education.[307] Other studies suggest that occupational licensing can hurt the broader economy,[308] particularly by decreasing consumer surplus and occupational mobility. However, studies examining the impact of occupational licensing on previously unregulated health care industries found that the quality of service improved.[309]

The historical context of professional licensure schemes serves as a stark reminder of the potential for abuse and discrimination. One account is that the introduction of licensure requirements in medicine, cosmetology, and plumbing combined with racist admission policies by unions and professional schools to dramatically decrease representation of African Americans in these disciplines.[310] Although these overtly racist mechanisms may be less prevalent today, the potential distributive impacts of licensure or accreditation schemes for machine learning practitioners shouldn't be ignored. Given the lack of representativeness on racial and gender dimensions in the field,[311] any policy likely to give preference to incumbent institutions and actors may reinforce or exacerbate these disparities.

---

[306] *See e.g.,* Josh Zumbrum, Occupational Licenses May Be Bad for the Economy, But Good for Workers Who Have Them, WALL STREET J. (Apr. 18, 2016), https://www.wsj.com/articles/BL-REB-35504; Morris M. Kleiner & Evan J. Soltas, A Welfare Analysis of Occupational Licensing in U.S. States, FED. RES. BANK MINNEAPOLIS (2019), https://www.oecd.org/economy/reform/welfare-effect-of-occup-licensing_Morris-Kleiner.pdf; Beth Redbird, The New Closed Shop? The Economic and Structural Effects of Occupational Licensure, 82 AM. SOCIO. ASS'N 3 (2017).

[307] DEPT. OF TREASURY OFF. ECON. POL'Y, COUNCIL OF ECON. ADVISERS & DEP'T OF LABOR, OCCUPATIONAL LICENSING: A FRAMEWORK FOR POLICYMAKERS 4 (2015), https://obamawhitehouse.archives.gov/sites/default/files/docs/licensing_report_final_nonembargo.pdf.

[308] Zumbrum, *supra* note 352; Kleiner & Soltas, *supra* note 352; Peter Q. Blair & Mischa Fisher, Does Occupational Licensing Reduce Value Creation on Digital Platforms?, NAT'L BUR. ECON. RSCH., Working paper No. 30388 (2022), https://www.nber.org/system/files/working_papers/w30388/w30388.pdf.

[309] D. Mark Anderson et al., *The Effect of Occupational Licensing on Consumer Welfare: Early Midwifery Laws and Maternal Mortality*, NAT'L BUR. ECON. RSCH. (Working Paper No. 22456, 2016), https://www.nber.org/papers/w22456; Marc T. Law & Sukkoo Kim, *Specialization and Regulation: The Rise of Professionals and the Emergence of Occupational Licensing Regulation*, 65 J. ECON. HIST. 3, 723 (2005), https://www.jstor.org/stable/3875015.

[310] David E. Bernstein, *Licensing Laws: A Historical Example of the Use Of Government Regulatory Power Against African-Americans*, 31 SAN DIEGO L. REV. 89 (1994).

[311] DANIEL ZHANG ET AL., THE AI INDEX 2021 ANNUAL REPORT 139–146 (2021), https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf.

Last, licensing regimes are particularly susceptible to capture. For example, research suggests that lobbying by physician interest groups is linked to a higher probability that a state will have occupational licensing in the healthcare industry.[312] The potential for special interest groups to have outsized impact on AI licensing regimes is particularly worrisome given licensing may make the frontier of AI technology inaccessible to most. While licensing can provide health and safety protections and improve the quality of services, the requirements can function as a barrier to entry in practice—particularly when the licensing requirements are not closely tied to occupational demands.[313]

VI. Auditing

Federal and state lawmakers, industry, and civil society organizations have all increasingly proposed AI audit requirements in response to rising concerns about the proliferation of unaccountable, biased, and otherwise harmful AI systems.[314] The chief agency responsible for advising the President on telecommunications and information policy[315] received 1,447 comments from the public in response to a request for information about AI audits and other AI accountability policies.[316] The CEO of OpenAI recently called upon Congress to require independent AI audits to ensure compliance with safety standards.[317] But implementation of one of the first AI audit laws in the United States—New York City's landmark bill requiring bias audits of AI used in hiring decisions—offers a glimpse into the technical and institutional feasibility challenges posed by AI auditing.

AI audits are generally understood as mechanisms for verifying that an AI system performs as is claimed and for evaluating an AI system's compliance with regulations or industry

[312] Benjamin J. McMichael, The Demand for Healthcare Regulation: The Effect of Political Spending on Occupational Licensing Laws, 84 SOUTHERN ECON. J. 1, 297.
[313] DEPT. OF TREASURY OFF. ECON. POL'Y, COUNCIL OF ECON. ADVISERS & DEP'T OF LABOR, *supra* note 307, at 4, 7.
[314] *See, e.g.*, Kate Kaye, *This Senate bill would force companies to audit AI used for housing and loans*, PROTOCOL (Feb. 8, 2022), https://www.protocol.com/enterprise/revised-algorithmic-accountability-bill-ai. State lawmakers have also proposed mandatory AI audits. *See, e.g.*, A4909, 220th Leg., Reg. Sess. (N.J. 2022); Stop Discrimination by Algorithms Act of 2023, B114, 25th Council (D.C. 2023).
[315] *About NTIA*, NTIA, https://www.ntia.gov/page/about-ntia (last visited Aug. 31, 2023). States are also proposing or passing less formalized impact and risk assessments, *see* H. 114 (Vt. 2023); S.B. No. 1103, *supra* note 79; H. 1974, 193d General Ct. (Mass. 2023); AB 331 (Ca. 2023) (proposing that developers and deployers of automated decision tools complete and document impact assessments that are submitted to the California Civil Rights Department).
[316] Cat Zakrzewski, Biden administration is trying to figure out how to audit AI, Wash. Post (Apr. 11, 2023), https://www.washingtonpost.com/technology/2023/04/11/biden-commerce-department-ai-rules/; Press Release, NTIA, NTIA Receives More Than 1,400 Comments on AI Accountability Policy (June 16, 2023), https://www.ntia.gov/press-release/2023/ntia-receives-more-1400-comments-ai-accountability-policy.
[317] *Hearing on Rules for AI*, *supra* note 263.

standards, where such exist.[318] Cited in the Trustworthy AI glossary published by NIST, the U.S. agency responsible for standard-setting, the Institute of Electrical and Electronics Engineers (the "IEEE") defines an audit, in its software engineering vocabulary standard, as a "systematic, independent, documented process for obtaining records, statements of fact, or other relevant information and assessing them objectively, to determine the extent to which specified requirements are fulfilled."[319]

In comparison to often less formalized impact or risk assessments,[320] a critical source of legitimacy in auditing is derived from the application of uniform accounting standards, which foster confidence in the consistency of evaluations and results.[321] These standards can focus on substance or process. For example, in financial accounting, there are two sets of standards: reporting standards that guide how financial information is to be reported to shareholders (e.g., instructing firms on when to recognize revenue, what is considered a liability or asset) and auditing standards that guide the auditor's role in verifying the information (e.g., how audit procedures should be supervised). Reporting standards are established by the Financial Accounting Standards Board, a U.S.-based standard-setting organization ("SSO"),[322] while auditing standards for public companies are established by the Public Company Accounting Oversight Board ("PCAOB"), a nonprofit corporation that is overseen by the U.S. Securities and Exchange Commission ("SEC").[323] Although reporting and auditing standards are commonly grouped together in many discussions of

[318] Marietje Schaake & Jack Clark, *Stanford Launches AI Audit Challenge*, Stanford Inst. Of Human-Centered A.I. (July 11, 2022), https://hai.stanford.edu/news/stanford-launches-ai-audit-challenge; Inioluwa Deborah Raji et al., *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, 2022 AAAI/ACM CONF. ON AI, ETHICS, & SOCIETY (2022), https://arxiv.org/pdf/2206.04737.pdf.

[319] TRUSTWORTHY & RESPONSIBLE AI RES. CTR., NAT'L INST. STANDARDS & TECHS., THE LANGUAGE OF TRUSTWORTHY AI: AN IN-DEPTH GLOSSARY OF TERMS, https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary (last visited Aug. 31, 2023); IEEE, ISO/IEC/IEEE International Standard - Systems and software engineering--Vocabulary, ISO/IEC/IEEE 24765-201 at 36 (2010), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8016712. For other definitions, *see Glossary of Computer System Software Development Terminology*, U.S. FOOD & DRUG ADMIN. (Nov. 6, 2014), https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-guides/glossary-computer-system-software-development-terminology-895 (describing ANSI's definition of an audit as "conduct[ing] an independent review and examination of system records and activities in order to test the adequacy and effectiveness of data security and data integrity procedures, to ensure compliance with established policy and operational procedures, and to recommend any necessary changes.").

[320] For example proposals, see *supra* note 315.

[321] *See* Patrick Hall, *What We Learned Auditing Sophisticated AI for Bias*, O'REILLY (Oct. 18, 2022), https://www.oreilly.com/radar/what-we-learned-auditing-sophisticated-ai-for-bias/; Ellen P. Goodman & Julia Trehu, *AI Audit-Washing and Accountability*, GERMAN MARSHALL FUND (Nov. 15, 2022), https://www.gmfus.org/news/ai-audit-washing-and-accountability.

[322] FASB establishes accounting and reporting standards for institutions following GAAP, see *About Us*, FASB, https://www.fasb.org/about (last visited Aug. 31, 2023).

[323] *See Auditing Standards*, PCAOB, https://pcaobus.org/oversight/standards/auditing-standards (last visited Aug. 31, 2023).

AI audit regulation, proposals encompass numerous notions of AI audits, with differences not only in the auditing process, including the use of uniform standards, but also in the parties conducting and reviewing the audits.

An "AI audit" or "algorithmic audit," as currently discussed within the AI and policy communities, carries several meanings.[324] AI audits can refer to internal audits primarily focused on model governance and risk management. Such internal audits draw upon robust literature about internal compliance programs, particularly in the financial services space, where audit teams distinct from business units validate models and assess the overall effectiveness of model risk management frameworks, including by assessing documented policies.[325] Alternatively, AI audits may refer to external audits similar to the financial accounting audits required for public companies under the nation's securities laws on an annual basis[326] or the FDA's routine audits of clinical trials to confirm a company's reported findings used in drug approval applications.[327]

The party conducting and reviewing the audit is also a key distinction between different types of audits. First-party audits, also referred to as internal audits, are conducted on a company's own AI system by auditors employed by the company.[328] In second-party audits, a customer or an entity contracted by the customer audits a business partner such as a supplier.[329] Because second-party audits can influence business or government decisions, these audits tend to be more formal than first-party audits.[330] For example, a government agency or company may audit an AI tool it bought, or is seeking to buy, from a third-party vendor. Third-party audits are conducted by parties that are supposed to be independent.[331] Borrowing terminology from the financial accounting space, a party is only independent if it receives no other remunerations from a company whose AI system is audited other than

[324] Reva Schwartz et al., NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence 45 (2022), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.
[325] *See, e.g.*, *id.*; BD. OF GOVERNORS OF THE FED. RSRV. SYS. & OFF. OF THE COMPTROLLER OF THE CURRENCY, SUPERVISORY GUIDANCE ON MODEL RISK MANAGEMENT 18 (2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf; OFFICE OF THE COMPTROLLER OF THE CURRENCY, MODEL RISK MANAGEMENT HANDBOOK 19–21, 84 (1st ed. 2021), https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.
[326] Securities Act of 1933, 15 U.S.C §§ 77a–77aa; Securities and Exchange Act of 1934, 15 U.S.C §§ 78a–78rr.
[327] Raji et al., *supra* note 318, at 16 tbl. 2.
[328] RYAN CARRIER & SHEA BROWN, TAXONOMY: AI AUDIT, ASSURANCE, & ASSESSMENT 4 (2021), https://forhumanity.center/web/wp-content/uploads/2021/09/ForHumanity.center_Taxonomy_AI_Audit_Assurance_Assessment.pdf; *What is Auditing?*, AM. SOC'Y FOR QUALITY (last visited Aug. 31, 2023), https://asq.org/quality-resources/auditing.
[329] Raji et al., *supra* note 318, at 2; *What is Auditing?*, *supra* note 328.
[330] *What is Auditing?*, *supra* note 328.
[331] Raji et al., *supra* note 318.

audit fees.[332] An even stronger notion of independence would require no remuneration, as happens with public inspections.[333] Oversight over audits can also be internal or external, with the former conducted by stakeholders employed or contracted by the company whose AI system is audited and the latter conducted by third-parties without such a relationship. Importantly, the third-party oversight can be provided by government agencies or public-interest institutions as well as private sector entities.[334]

A. Technical Feasibility: Identifying Uniform and Administrable Evaluation Criteria can be Difficult

AI audits suffer from a number of technical feasibility constraints. First, there is a significant gap between the types of values and AI principles regulators envision audits measuring (e.g., privacy, robustness, or transparency), and the existing methods for evaluating those values and principles in AI systems. Second, the sophistication of AI systems and their integration into complex software systems can make audit execution intractable.

On the first, effective AI audits will require standards that establish uniform interpretations of the characteristics of the audited AI system. High-level proposals to audit for adherence to broad principles can be too difficult to put into practice, let alone implement in a consistent manner throughout an industry. Conversely, audits that focus too narrowly or only on specific metrics may prevent evaluations that capture the full scope of concerning practices or behaviors.[335] For example, an AI audit focused on fairness that requires a system "does not discriminate" will likely be interpreted in very different ways while mandating the monitoring of only one specific fairness metric may fail to rein in AI systems that are biased in different ways[336] or, in the worst case, even exacerbate disparities by

[332] Carrier & Brown, *supra* note 328, at 4 (citing Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (2002)). For another discussion of independence, *see* Jacob Metcalf, Ranit Singh, Emmanuel Moss, & Elizabeth Anne Watkins, *Witnessing Algorithms at Work: Toward a Typology of Audits*, DATA & SOCIETY (Aug. 11, 2022), https://points.datasociety.net/witnessing-algorithms-at-work-toward-a-typology-of-audits-efd224678b49.
[333] Esther Duflo et al., *Truthtelling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*, 128 Q. J. ECON. 1499 (2013).
[334] Raji et al., *supra* note 318, at 2.
[335] For a discussion of AI audits focused on fairness and transparency, *see, e.g.*, Shea Brown et al., *The Algorithm Audit: Scoring the Algorithms That Score Us*, 8 BIG DATA & SOCIETY, January–June 2021, at 1 (2021), https://journals.sagepub.com/doi/full/10.1177/2053951720983865.
[336] Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1 (2019).

focusing efforts on an inappropriate metric which may be statistically incompatible with more relevant conceptualizations of fairness in a given context.[337]

New York City's experience with its hiring law illustrates how a legal requirement to audit, absent standards, can be a challenging feat. Originally slated to take effect in January 2023,[338] NYC twice delayed enforcement because of the high volume of public comments and requests for clarification about the audit requirements.[339] The final rule, published in April 2023, clarifies the bias audit's required metrics (e.g., "impact ratio" by sex, race/ethnicity, and intersectional categories) and other information, such as when a company is exempted from the requirement to conduct the bias audit using its own historical data.[340] But disagreement over the exact contours of the final rule still remains, as does uncertainty about various requirements, such as the required labeling of training and testing data.

Literature outside of the AI context points to the benefits of standards to ameliorate these challenges. For example, rules-based financial audits in Belgium decreased errors and increased the independence of auditors.[341] But uniform standards do not spring up overnight. Though financial audits date back to the mid-19th century, financial accounting in the United States was not standardized until the 20th century, when financial regulators mandated financial audits for public companies in response to the 1929 stock market crash.[342]

Policymakers are increasingly turning to SSOs in hopes that they can define key AI terms and practices. SSOs bring technical expertise across industry together to build consensus

---

[337] *See e.g.*, Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, PROC. 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 797 (2017); Kleinberg et al., *supra* note 70.

[338] Richard Vanderford, *New York's Landmark AI Bias Law Prompts Uncertainty*, WALL ST. J. (Sep. 21, 2022, 5:30 AM), https://www.wsj.com/articles/new-yorks-landmark-ai-bias-law-prompts-uncertainty-11663752602.

[339] Lindsay Stone, *NYC Issues Final Regulations for Automated Employment Decision Tools Law, Delays Enforcement to July 5, 2023*, JDSUPRA (Apr. 13, 2023), https://www.jdsupra.com/legalnews/nyc-issues-final-regulations-for-3612453/.

[340] *Id*.

[341] Joseph V. Carcello et al., *Rules Rather Than Discretion in Audit Standards: Going-Concern Opinions in Belgium*, 84 ACCT. REV. 1395 (2009).

[342] Goodman & Trehu, *supra* note 321; Thomas Bourveau et al., *Public Company Auditing Around the Securities Exchange Act* (Columbia Business School Research Paper, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3837593. The SEC did not establish the first accounting standard setting organization until 1938-39. See Stephen Zeff, "Evolution of US Generally Accepted Accounting Principles ('GAAP')", working paper, https://www.iasplus.com/en/binary/resource/0407zeffusgaap.pdf

around common guidelines, definitions, and rules for certain technologies.[343] Technical standards, particularly those issued by NIST and by international SSOs like International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), and IEEE have been critical to advancing interoperability and uniformity in many other technical sectors.[344] For example, compliance with cybersecurity standards promulgated by NIST and international bodies like ISO/IEC has become industry norm, helping certify that vendors and companies implement baseline practices to protect data and systems.[345]

Using AI standards set by SSOs could similarly provide confidence that AI audits consistently verify an AI system is of a minimum quality. The European Commission has embraced this hope, hitching critical aspects of the EU AI Act on the ability of SSOs like the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) to develop such standards.[346] The EU AI Act requires that third-parties assess whether high-risk AI systems conform with "harmonised standards" set by CEN and CENELEC, establishing what some have argued is a de facto auditing requirement.[347]

But SSOs are far from reaching consensus on many AI-related reporting standards. Many key terms used by those promoting trustworthy AI (e.g., "bias") are defined abstractly.[348] And even where the SSOs have defined some metrics to measure bias,[349] there is no consensus on what an AI audit focused on mitigating bias should focus on. Furthermore,

---

[343] See e.g., *Introducing Standards*, AI STANDARDS HUB, https://aistandardshub.org/resource/main-training-page-example/1-what-are-standards/ (last visited Aug. 24, 2023); NAT'L SEC. COMM'N ON A.I., INTERIM REPORT AND THIRD QUARTER RECOMMENDATIONS 206 (2020), https://www.nscai.gov/wp-content/uploads/2021/01/NSCAI-Interim-Report-and-Third-Quarter-Recommendations.pdf.
[344] *See* NAT'L SEC. COMM'N ON A.I., INTERIM REPORT AND THIRD QUARTER RECOMMENDATIONS, *supra* note 343, at 205.
[345] Alladean Chidukwani et al., *A Survey on the Cyber Security of Small-to-Medium Businesses: Challenges, Research Focus, and Recommendations*, 10 IEEE ACCESS 85701, 85702 (2022).
[346] CEN and CENELEC are two regional standard-setting bodies—private, independent nonprofits that shepherd across the 34 European country members the setting of technical standards. Clément Perarnaud, *With the AI Act, We Need To Mind the Standards Gap*, CTR. FOR EUR. POL'Y STUD., https://www.ceps.eu/with-the-ai-act-we-need-to-mind-the-standards-gap/ (last visited Aug. 24, 2023); Hadrien Pouget, *Standard Setting*, AI ACT NEWSLETTER, https://artificialintelligenceact.eu/standard-setting/ (last visited Aug. 31, 2023). The Digital Platforms Commission Act of 2023, proposed Senator Michael Bennett in May 2023, also includes a focus on technical standards, proposing a "Federal Digital Platform Commission" consider establishing technical standards including on data portability, interoperability, and age verification. Digital Platforms Commission Act, S. 1671, 118th Cong. (2023).
[347] *See, e.g.*, Jakob Mökander et al., *Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation*, 32 MINDS & MACHINES 241 (2021).
[348] *See, e.g.*, *Information Technology — Artificial Intelligence (AI) — Bias in AI Systems and AI Aided Decision Making*, INT'L ORG. FOR STANDARDIZATION (2021), https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:24027:ed-1:v1:en, (defining "bias" as "systematic difference in treatment of certain objects, people, or groups in comparison to others").
[349] *Id.*

assessing an AI system's realization of each trustworthy AI principle (e.g., fairness, privacy-preserving, accuracy) necessitates that a company monitors, and an auditor verifies, completely different qualitative or quantitative metrics. And the technical feasibility of calculating each of these metrics varies because they require a company to maintain different data and information, internal governance procedures, and documentation.

Although the ostensible neutrality and transparency of SSOs engenders trust in their standards, the process of setting standards can be quite time-consuming and laborious as technical committees require vast amounts of research to support the standard and meet several times, sometimes over years, to reach consensus.[350] Biometric standards provide a useful comparison, as policymakers flocked to biometric identifiers to increase airport security in the wake of 9/11.[351] But research on the technology had began decades earlier. For example, NIST and the FBI began researching technologies for automated fingerprint matching in 1967 with a standard on fingerprint ridges published in 1986 and a standard that would enable interoperability of automated fingerprint live scans in 1993.[352] Despite decades of research and standard-setting, 10-fingerprint collection at all visa-issuing posts and U.S. airports did not begin until 2008, after NIST conducted years of research on fingerprint testing and published multiple standards on fingerprinting and biometrics.[353]

Consensus standards on AI may take similar time and research investments. ISO and the IEC have been working since 2017 on a variety of AI-related standards through its joint task force sub-committee on AI.[354] IEEE has similarly spent years on AI standards particularly related to ethics.[355] Even NIST's standards, which do not require international agreement, can take years to develop as the evidence base for the standards is built up and verified. The significant cost borne by private industry involved in standard-setting is only likely to exacerbate these challenges. Participating in meetings is expensive with estimates

---

[350] Matt Sheehan et al., *What Washington Gets Wrong About China and Technical Standards*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Feb. 27, 2023), https://carnegieendowment.org/2023/02/27/what-washington-gets-wrong-about-china-and-technical-standards-pub-89110.
[351] *E.g.*, Nat'l Comm'n on Terrorist Attacks, *The 9/11 Commission Report* 386 (2004), https://www.9-11commission.gov/report/911Report.pdf; Neal Latta, *US-VISIT Biometrics Overview*, DEP'T OF HOMELAND SEC., https://www.nist.gov/system/files/documents/2021/03/05/ansi-nist_archived_2007_workshop1_latta-visit-overview.pdf.
[352] Nat'l Sci. & Tech. Council, *Biometrics in Government Post-9/11* 43 (Aug. 2008), https://irp.fas.org/eprint/biometrics.pdf; Kenneth R. Moss, Chapter 6: Automated-Fingerprint Identification System (AIFS), *in* THE FINGERPRINT SOURCEBOOK 6-4, 6-16, 6-17 (U.S. Dep't of Justice, 2011) https://www.ojp.gov/pdffiles1/nij/225326.pdf.
[353] *Id*. at 9–10.
[354] *ISO/IEC JTC 1 SC 42 — Artificial Intelligence,* https://www.iso.org/committee/6794475.html (last visited Sep. 1, 2023).
[355] *See e.g., The IEEE Global Initiative on Ethics of Autonomous and Intelligence Systems*, IEEE, https://standards.ieee.org/industry-connections/ec/autonomous-systems/ (last visited Sep. 1, 2023).

that it can cost a company over $300,000 per year to ensure one standards engineer participates.[356]

The speed of AI innovation may further complicate standard-setting as standards become obsolete, perhaps at a greater rate than prior technologies such as for fingerprinting. For example, a watermarking standard might be state-of-the art today but quickly become obsolete in the future. SSOs may then choose to focus on only rudimentary standards more likely to withstand changes in technology, but this may limit the standard's utility. Another option is to establish programs, such as the SOC-2 certification in cybersecurity, that verify not whether a company adheres to specific technical standards, but whether it has established and complies with its own rigorous internal controls.[357] Such an approach could be far more adaptable. Standards created in a less formalized fashion–e.g., by industry in-house–would be more able to adapt to changing technology but are also more susceptible to industry capture.

AI audits may also be technically infeasible where the targeted system is a platform technology or requires continuous updating. Discrete AI systems, for example an AI tool used for hiring or credit decisions, may be well suited to auditing focused on ensuring the system is trustworthy, accurate, and reliable. However, audits of all AI or ML could require a company providing a platform service, for example a webpage or streaming service, to audit dozens of algorithms that run in parallel. Auditors could struggle to isolate algorithms or expend significant resources auditing all the algorithms on the larger platform even where the actual intent of the audit is to evaluate the system as a whole. Similarly, requirements for audits whenever an AI system is updated might become unwieldy where companies make minor, routine adjustments. In some cases, this could disincentivize desirable speedy updates. For example, in the wake of Christchurch, Australia passed a law requiring live streaming, video sharing platforms, and other content sharing services to remove access to "abhorrent" material within a "reasonable" amount of time (although the initial proposal required action within one hour).[358] Compliance with such a law would in

[356] Sheehan et al., *supra* note 350.
[357] For a high-level overview on the pros and cons of SOC-2-style certification, see Thomas Ptacek, *SOC2: The Screenshots Will Continue Until Security Improves*, FLY.IO BLOG (July 18, 2022), https://fly.io/blog/soc2-the-screenshots-will-continue-until-security-improves/.
[358] *Abhorrent Violent Material Act Fact Sheet*, AUSTRALIAN GOVERNMENT — ATTORNEY-GENERAL'S DEPARTMENT (July 16, 2019), https://www.ag.gov.au/crime/publications/abhorrent-violent-material-act-fact-sheet;
https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id:%22legislation/bills/s1201_aspassed/0000%22; *see also* Jonathan Shieber, *Australia passes law to hold social media companies responsible for "abhorrent violent material*, TECHCRUNCH (Apr. 4, 2019, 6:11 PM), https://techcrunch.com/2019/04/04/australia-passes-law-to-hold-social-media-companies-responsible-for-abhorrent-violent-material/; Ry Crozier, *Australia's 'world-first' social media laws could require action*

many cases require updating algorithms used to identify and promote content. Thus, AI audit requirements could benefit from careful scoping to specific use cases or discrete AI systems and avoid new audits after any and all updates.

B. Institutional Feasibility: The Importance of Maintaining Auditor Independence

The institutional design of an AI auditing regime can make or break the effectiveness of such audits, even where the goal, standards, and methodology are defined. Under-defined standards, particularly in comparison to bright-line rules, are at risk of inconsistent implementation, especially by insufficiently trained auditors. For example, even when observing identical conditions, inspectors for health code violations disagreed 60 percent of the time on whether to cite a major violation.[359] The accuracy and utility of audits are also severely undermined when auditors are not independent or are denied robust access to information about the company or system audited.[360] Auditing programs with private sector auditors are difficult to design and implement with sufficient independence and professionalism, but programs that rely upon public sector auditors can quickly become limitless mandates unmanageable by agencies often under-resourced and under-staffed.

Audits conducted by third-parties with minimal conflicts of interests and independence from the company being audited are the most reliable.[361] Robust literature demonstrates this across a variety of sectors: Audits are more accurate where the auditor cannot cross-sell non-auditing services to, is not paid or chosen by, and has a lesser degree of familiarity (i.e., does not have a close relationship established through repeat interactions) with the company being audited.[362] For example, randomized controlled trials have demonstrated that environmental third-party audits are more truthful when the auditors are paid through government funding instead of the company being audited.[363]

The virtues of completely independent audits have perhaps motivated the calls for the FTC or a new government entity, such as a "Federal Digital Platforms Commission", to enforce

---

within an hour, ITNEWS (Apr. 4, 2019, 12:52 PM), https://www.itnews.com.au/news/australias-world-first-social-media-laws-could-require-action-within-an-hour-523389.

[359] Daniel E. Ho, *Does Peer Review Work? An Experiment of Experimentalism*, 69 STAN. L. REV. 1 (2017) [hereinafter *Does Peer Review Work*?].

[360] *See, e.g.*, Duflo et al., *supra* note 333; Veronica Toffolutti et al., *Evidence points to 'gaming' at hospitals subject to National Health Service Cleanliness Inspections*, 36 HEALTH AFFS. 355 (2017).

[361] Raji et al., *supra* note 318.

[362] Monika Causholli et al., *Future Nonaudit Service Fees and Audit Quality*, 31 CONTEMP. ACCT. RSCH. 681 (2014).

[363] Duflo et al., *supra* note 333.

AI audits requirements.[364] Absent significant changes in the AI workforce and the pace of AI innovation, such proposals are unrealistic. Depending on the breadth of AI systems subject to these audits, a federal regulator could have an insurmountable volume of AI systems to audit. In addition to perhaps being technically infeasible, as explained above, this task would be institutionally infeasible. Auditing or reviewing large volumes of audits would be difficult enough for an agency already well-versed in both AI and scrutinizing the private sector. The FTC, for instance, is building AI expertise[365] and has deep experience investigating potential legal violations to bring enforcement actions, but it currently lacks the technical and institutional capacity necessary to run a full-scale AI auditing program. Given that Congress may be hesitant to further empower an agency it has previously defunded for overstepping its mandate,[366] it appears unlikely that the FTC would receive the necessary authority and appropriations to build that capacity. Even if it did, the technical talent gap facing the federal government would likely pose an insurmountable barrier to the effective administration of such a program in the near term.[367]

Relying solely on the private sector, however, also faces serious institutional challenges. Here, the NYC hiring law is again instructive. It requires "independent auditors" that are "capable of exercising objective and impartial judgment" and have not used, developed, or distributed the AI system, been employed by the company being audited, or have a "direct financial interest or a material indirect financial interest" in the company being audited or vendor of the AI system.[368] This explicitly precludes first-party and second-party audits conducted internally. Companies subject to the requirement could rely upon a cottage industry of AI auditing companies that has cropped up in response to auditing proposals (or perhaps has identified a business opportunity and successfully convinced policymakers

[364] *See, e.g.*, Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. (2022) (proposing the FTC require "covered entities" "perform impact assessments . . . including through participatory design, independent auditing . . . ."); Digital Platforms Commission Act of 2023, *supra* note 346 (proposing the establishment of a Federal Digital Platforms Commission that establishes requirements for "auditing, accountability, and explainability of algorithmic processes" and establishes "transparency and disclosure obligations" for "systematically important platforms" that enables "third-party audits to ensure the accuracy of any public risk assessments required").

[365] Samuel Levine, Director, Bureau of Consumer Protection, Fed. Trade Comm'n, Believing in the FTC (Apr. 1, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/Remarks-to-JOLT-4-1-2023.pdf at 9.

[366] Congress allowed the FTC's funding to lapse in the wake of the "Kid-Vid controversy." J. Howard Beales, III, *Advertising to Kids and the FTC: A Regulatory Retrospective That Advises the Present* 7, FED. TRADE COMM'N (Mar. 2, 2004), https://www.ftc.gov/sites/default/files/documents/public_statements/advertising-kids-and-ftc-regulatory-retrospective-advises-present/040802adstokids.pdf; *see also* Tracy Westen, *Government Regulation of Food Marketing to Children: The Federal Trade Commission and the Kid-Vid Controversy*, 39 LOY. L.A. L. REV. 79 (2006).

[367] *See* MASLEJ ET AL., *supra* note 99, at 245 fig. 5.1.9 (2023) (fewer than 1% of new A.I. PhDs work in government).

[368] N.Y.C. Dep't of Consumer & Worker Prot., *supra* note 85..

of their merits),[369] but academic literature questions whether company-selected third-party auditors can ever be fully independent.[370]

Effective third-party audits require auditors to have access to the AI system and company data, records, and documentation to conduct accurate and consistent audits,[371] but companies may severely limit an auditor's access and influence an auditor's inquiry. For example, companies can thwart independent auditing by requiring pre-publication review of an audit, invoking trade secret protection and requiring NDAs, or obscuring access to the service including through paywalls and prohibitive terms of service.[372] HireVue, a large vendor of AI hiring software, publicized its software as having passed a civil rights audit. In reality, HireVue appears to have severely limited the scope of the "audit" conducted by O'Neil Risk Consulting and Algorithmic Auditing (ORCAA) and carefully controlled the messaging about it,[373] only allowing access to their audit after signing an NDA.[374] Pymetrics also claimed to have a "neutral third party" audit of its AI hiring tool.[375] But through a so-called "collaborative audit," Pymetrics framed the questions that the auditors asked, rendering the exercise far from independent.[376]

The HireVue and Pymetrics examples illustrate broader worries that AI audits are more a ploy for positive media attention than genuine efforts to evaluate an AI system's fairness, accuracy, and robustness.[377] Such concerns are not assuaged by the origin story of the NYC hiring law. Pymetrics created an open audit tool and then worked with the political strategy firm Tusk Strategies to lobby for the passage of the NYC bill, including by securing seven cosponsors, building a "network of grassroots partners who could provide third-party validation for the bill with legislators in the form of meetings and testimony," undertaking

[369] Kate Kaye, *A New Wave of AI Auditing Startups Wants to Prove Responsibility Can Be Profitable*, PROTOCOL (Jan. 3, 2022), https://www.protocol.com/enterprise/ai-audit-2022.
[370] Raji et al., *supra* note 318.
[371] *Id.*; Goodman & Trehu, *supra* note 321.
[372] Raji et al., *supra* note 318, at 7.
[373] *Id.*
[374] Hilke Schellmann, *Auditors Are Testing Hiring Algorithms for Bias, But There's No Easy Fix*, MIT TECHNOLOGY REVIEW (Feb. 11, 2021), https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/.
[375] "'What Pymetrics is doing, which is bringing in a neutral third party to audit, is a really good direction in which to be moving,' says Pauline Kim, a law professor at Washington University in St. Louis, who has expertise in employment law and artificial intelligence. 'If they can push the industry to be more transparent, that's a really positive step forward.'" *Id.*
[376] Raji et al., *supra* note 318, at 7.
[377] Goodman & Trehu, *supra* note 321.

an aggressive PR campaign, and ensuring Pymetrics's "legislative efforts [were] recognized by Fast Company as a finalist for their 2021 World Changing Ideas awards."[378]

Effective third-party audits also require auditors to receive necessary training and expertise, to conduct accurate and consistent audits.[379] In the health inspection context, a randomized trial showed that accuracy and consistency improved with increased training and peer review.[380] But it may also take significant time and resources to professionalize the AI auditing community.[381] Financial accounting audits again provide a useful comparison: It took several decades before financial accountants started to professionalize, and even after the post-1929 stock market crash professionalization, self-regulation proved insufficient in preventing the Enron financial scandal.[382] Numerous questions about auditor independence, access to information, and professionalism and post-audit actions[383] thus implicate institutional feasibility concerns of AI audits.

Regulatory oversight can make auditing regimes more independent, trustworthy, and accurate. One option would be to task an entity like the PCAOB with oversight of AI auditors. The PCAOB, a five-member nonprofit board established by the Sarbanes-Oxley Act in the wake of the Enron scandal and subject to SEC oversight, has a joint mission of promulgating auditing standards for the financial accounting industry, and providing oversight to ensure that those standards are followed.[384] Accounting firms are required to register with the PCAOB in order to provide certain professional services. By registering with the PCAOB, all accounting firms agree to follow PCAOB auditing standards on the audits regulated by the entity, and to submit to PCAOB oversight. The PCAOB's oversight mechanism primarily consists of inspections of the audits performed by registered accounting firms.

---

[378] Khari Johnson, *Pymetrics Open-sources Audit AI, An Algorithm Bias Detection Tool*, VENTUREBEAT (May 31, 2018 3:47PM), https://venturebeat.com/ai/pymetrics-open-sources-audit-ai-an-algorithm-bias-detection-tool/; Matt O'Brien, *NYC Aims to be First to Rein in AI Hiring Tools*, AP NEWS (Nov. 19 2021 5:11AM), https://apnews.com/article/technology-business-race-and-ethnicity-racial-injustice-artificial-intelligence-2fe8d3ef7008d299d9d810f0c0f7905d; *Enacting First-Mover AI Legislation*, TUSK STRATEGIES, https://tuskstrategies.com/wins/enacting-first-mover-ai-legislation/ (last visited Sep. 1, 2023).
[379] Raji et al., *supra* note 318; Goodman & Trehu, *supra* note 321.
[380] *Does Peer Review Work?*, *supra* note 359.
[381] Raji et al., *supra* note 318.
[382] Goodman & Trehu, *supra* note 321.
[383] Raji et al., supra note 318.
[384] *About*, PCAOB, https://pcaobus.org/about.

Academic research has found evidence that PCAOB inspections have improved audit quality[385]—both in the U.S. and abroad.[386] Nonetheless, the PCAOB is an imperfect model. Some critics have accused the PCAOB of overreach and government waste.[387] And accounting firms subject to PCAOB oversight criticize the PCAOB for penalizing overly technical violations that, they argue, slow down the audit process without improving audit quality. Commentators on the other side have critiqued the PCAOB for being too deferential to the accounting firms it regulates.[388] Such critics commonly point to the high rate of deficiencies in audits inspected by the PCAOB–an expected 40% in 2022[389]–and question why the deficiency rate remains so high, suggesting that harsher penalties are needed. Furthermore, establishing an entity similar in expertise and size may be difficult: In 2022 alone, PCAOB set 30 audit standards, inspected over 207 audit firms, and reviewed over 800 audit engagements.[390]

C. Auditing's Tensions: Effective but Expensive

First, AI audits that prioritize certain values may create horizontal misalignment through direct conflict with the realization of other values. For example, auditing requirements focused on ensuring an AI system is privacy-preserving, including by following data minimization principles, may make it harder for those same AI systems to be assessed for bias.[391] Similar tradeoffs have been documented between bias and accuracy and accuracy and interpretability.[392]

Second, the technical and institutional challenges to establishing reporting standards for many key trustworthy AI principles highlights gaps in existing regulatory regimes and legal doctrine, particularly around the distribution of liabilities and burdens. In particular, the availability of commercial off-the-shelf AI systems raises questions about the proper

[385] Joseph V. Carcello et al., The Effect of PCAOB Inspections on Big 4 Audit Quality, 23 RSCH. IN ACCT. REG. 85 (2011).
[386] Phillip T. Lamoreaux, *Does PCAOB inspection access improve audit quality? An examination of foreign firms listed in the United States*, 61 J. ACCT. & ECON. 313 (2016). Research has found additional benefits of PCAOB oversight, such as greater reporting credibility. Brandon Gipper et al., *Public Oversight and Reporting Credibility: Evidence from the PCAOB Audit Inspection Regime*, 33 REV. FIN. STUD. 4532 (2020).
[387] *See, e.g.*, Hester M. Peirce, *PCAOB's Ballooning Budget*, US SEC. & EXCH. COMM'N (Dec. 23, 2022), https://www.sec.gov/news/statement/peirce-pcaob-budget-20221223.
[388] Daniel L. Goelzer, *Audit Oversight and Effectiveness*, CPA J. (Feb. 2021), https://www.cpajournal.com/2021/02/22/audit-oversight-and-effectiveness/.
[389] Press Release, PCAOB, PCAOB Report: Audits With Deficiencies Rose for Second Year In a Row to 40% in 2022 (July 25, 2023), https://pcaobus.org/news-events/news-releases/news-release-detail/pcaob-report-audits-with-deficiencies-rose-for-second-year-in-a-row-to-40-in-2022.
[390] *About*, PCAOB, *supra* note 384.
[391] Gupta et al., *supra* note 72.
[392] Giorgos Myrianthous, *Understanding the Accuracy-Interpretability Trade-Off*, TOWARDS DATA SCIENCE (Oct. 6, 2021), https://towardsdatascience.com/accuracy-interpretability-trade-off-8d055ed2e445.

allocation of liability between developers and deployers. In employment settings, liability typically resides with employers to ensure fair hiring practices. Consistent with this view, the NYC hiring algorithm audit law requires *employers* to audit the hiring tools they use, even if they did not develop the tool. Some disagree with this approach, instead arguing that the third-party vendors that develop and supply the AI tools should be held liable as they are best situated to ensure the AI tools do not discriminate.[393]

The NYC hiring law also exposes gaps in existing antidiscrimination law and is perhaps a reaction to the difficulty plaintiffs face in bringing successful disparate impact claims for algorithmic discrimination.[394] Supreme Court decisions have narrowed plaintiffs' ability to successfully challenge employers for the use of hiring practices that have a disproportionately adverse impact on a protected class.[395] A hiring algorithm, in comparison to an HR representative, is arguably harder for plaintiffs to interrogate. Even where a plaintiff can show a disparate impact, an employer that justifies the policy by showing a legitimate objective can shift the burden back to the plaintiff to prove there was a less discriminatory alternative that would achieve that same legitimate objective. Given the technical complexities of AI systems—not to mention the massive amounts of data and compute used by many companies with AI products and services—and the ability of companies to shield their AI systems from scrutiny (e.g., by claiming trade secrets), plaintiffs are likely to struggle to show a less discriminatory alternative, particularly a less discriminatory *algorithm*.

An audit requirement to ensure an AI system is not discriminating can thus be viewed as a way of shifting the burden to the employer. The NYC hiring law, for example, relies upon the EEOC's 80% rule to determine whether an AI hiring tool is discriminatory without addressing business necessity or less discriminatory alternatives. An AI audit could be seen to shift the burden to employers by, for example, requiring companies to audit and document potential less discriminatory alternatives. Current disparate impact doctrine places the burden of proving a less discriminatory alternative on plaintiffs. The call for AI

---

[393] *E.g.,* J. Edward Moreno, *Workplace AI Vendors, Employers Rush to Set Bias Auditing Bar*, BLOOMBERG L. (Mar. 13, 2023, 3:30 AM), https://news.bloomberglaw.com/daily-labor-report/workplace-ai-vendors-employers-rush-to-set-bias-auditing-bar; Roshan Abraham, *Business Lobby Tries to Weaken Law Regulating Bias in Hiring Algorithms*, Vice (Mar. 6, 2023), https://www.vice.com/en/article/n7ejn8/business-lobby-tries-to-weaken-law-regulating-bias-in-hiring-algorithms.

[394] For discussion of difficulties plaintiffs face bringing disparate impact claims, *see, e.g.*, DAVID H. CARPENTER, DISPARATE IMPACT CLAIMS UNDER THE FAIR HOUSING ACT, Cong. Rsch. Serv., R44203 (Sept. 24, 2015), at 2, https://crsreports.congress.gov/product/pdf/R/R44203; Joseph A. Seiner, *Plausibility and Disparate Impact*, 64 Hastings L. J. 287 (2013).

[395] *See id.*; *SCOTUS Sets High Bar For Those Bringing Race Discrimination Cases*, FISHER PHILLIPS (Mar. 31, 2020), https://www.fisherphillips.com/en/news-insights/scotus-sets-high-bar-for-those-bringing-race-discrimination-cases.html.

audits may hence illustrate the need for resolving deeper questions in the structure of employment discrimination law.

Third, AI audits can closely resemble requirements for disclosures, registration, and other regulatory regimes. Some proposals may be better characterized as transparency or disclosure requirements than as audits as they focus on simply requiring greater documentation and increasing the ability of the public or government to inspect and test an AI system.[396] Inspections by government agencies can also resemble third-party audits. For example, the FDA conducts "pre-approval inspections" to assess a drug manufacturing site's readiness for commercial manufacturing, verify the consistency of a drug application's description to the actual manufacturing methods etc., and to *audit* the data submitted in a drug application.[397] Audits that also require auditors receive certain training or accreditation can also resemble licensing.

Fourth, extensive audit requirements may necessitate extensive compliance regimes that asymmetrically burden certain industry players (e.g., small companies with limited resources or companies providing platform services with continuous updating). Particularly expansive or ill-defined audits may exacerbate these challenges as regulated entities and auditors may expend significant effort interpreting the requirement. Audits that focus on ensuring a company is complying with its own rigorous internal controls, rather than specific technical standards (e.g., SOC-2) are unlikely to alleviate this compliance burden.

VII. Discussion

With so much unknown about AI's risks or the full scope of its applications, a broad coalition in support of regulation appears to have emerged.[398] But the harms that animate these calls are vastly different in kind and degree—ranging from fears that discriminatory AI and deepfakes will undermine our democracy to concerns that AI-controlled weapons or AI-assisted bio-attacks could destroy humanity. Yet, it is infeasible—and sometimes impossible—to satisfy every goal of regulation. Each of the four categories of AI regulation we describe suffers from its own alignment problems. Some proposals may be technically and institutionally infeasible and fail to reduce targeted harms. Others may worsen the problems they intended to solve or introduce entirely new harms.

[396] Schwartz et al., *supra* note 324, at 45.
[397] *Pre-Approval Inspection (PAI): An Expert Guide to Preparation*, U.S. FOOD & DRUG ADMIN. (Jan. 18, 2022), https://www.thefdagroup.com/blog/pre-approval-inspection-pai-expert-guide-preparation. Denise DiGlulio, *FDA's Pre-Approval Inspection (PAI) Program and How to Prepare for a Successful Outcome*, U.S. FOOD & DRUG ADMIN. (2015), https://www.fda.gov/files/drugs/published/FDA%E2%80%99s-Pre-Approval-Inspection-(PAI)-Program-and-How-to-prepare-for-a-successful-outcome.pdf.
[398] *See, e.g.*, *supra* notes 5, 17, 316.

AI regulation cannot be "all things to all people."[399] Regulation will present real tradeoffs, and designing effective, enforceable schemes will require prioritizing specific goals over others. Achieving regulatory alignment and consensus on those goals will not be easy. But doing so will be essential to building an AI ecosystem that is safe, beneficial, and *effective* for all.

A. Misalignment in AI Regulation

AI regulation should be well-suited to achieving its intended goal or goals. Yet developing AI regulation that works effectively – particularly in light of competing concerns – is not easy. Reasonable people may disagree about what regulatory outcomes will improve Americans' lives and strengthen the country. But, at a minimum, the impacts of regulation and how regulation may require tradeoffs with other policy goals must be understood. Our analysis, however, reveals that neither attainment of the intended goal nor honest deliberation about tradeoffs are assured in the discourse about, or implementation of, four common AI regulation proposals. Misalignment is rampant across proposed regulation, with five common themes.

First, many kinds of AI regulation are beset by similar issues of technical and institutional feasibility. From a technical perspective, regulations that apply to a particular category of AI systems (e.g., LLMs more capable than GPT-4) may struggle to precisely articulate criteria for coverage. Compounding that difficulty, AI systems are frequently updated and modified for many purposes, including to fix vulnerabilities and improve accuracy for particular use-cases. Regulators will have to determine when such updates should trigger new legal obligations (e.g., re-registration or audits), balancing the goals of regulation against the benefits of quick updates, which may themselves mitigate many risks.

From an institutional perspective, enforcing AI regulations will require significant domain expertise, but government agencies face a daunting shortage of AI talent at present. That challenge is most acute for resource-intensive programs like an agency for government auditing or licensing AI, but any effort to enforce regulations across the highly decentralized and heterogeneous AI ecosystem will face similar issues. Policymakers must account for the AI talent gap in designing a robust regulatory regime, while also working to build public-sector AI expertise.

Second, proposals to regulate AI suffer from regulatory mismatch, with values—articulated in response to perceived or observed harms—vertically misaligned with

[399] Mark A. Lemley, *The Contradictions of Platform Regulation*, 1 J. FREE SPEECH L. 303, 335 (2021).

regulatory objectives, leading to unintended consequences. Often, technical and institutional challenges make the proposal's ability to achieve its goals infeasible. But the mismatch may also result from a proposal's misalignment with the harm it is intended to reduce.

Non-AI regulatory reform may better address a number of risks. Returning to the biosecurity example, manufacturing bioweapons is already illegal.[400] The MIT study of LLM-related biosecurity risks alludes to laboratories that are not in the International Gene Synthesis Consortium (IGSC) and which therefore may be willing to synthesize influenza strains.[401] Investigations of non-IGSC laboratories and audits of contractors for pathogens to ensure compliance with existing restrictions on manufacturing and distributing influenza strains may more effectively prevent bioweapons proliferation.[402] In considering whether AI-specific regulations are warranted in a particular context, policymakers should first ask: Are the harms being addressed specific to AI systems, or do they point instead to a non-AI regulatory solution?[403]

Third, specific regulatory interventions often place different values and goals of regulation in conflict, with such horizontal misalignment potentially necessitating tradeoffs.[404] For instance, speculative risk about the future destruction of humanity might ground demands to restrict open models, but concrete risks of bias may be more easily assessed and mitigated with transparency and open models. Ensuring a model is fully privacy-preserving, non-discriminatory, explainable, and accurate may not be technically achievable. AI regulatory proposals can fall into a trap by claiming to address all that ails

[400] The Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction, Mar. 26, 1975, 26 U.S.T. 583, 1015 U.N.T.S. 163.
[401] Soice et al., *supra* note 12, at 2.
[402] Victoria Sutton, *Emerging Biotechnologies and the 1972 Biological Weapons Convention: Can It Keep Up with the Biotechnology Revolution?*, 2 TEX. A&M L. REV. 695, 713 (2015) (noting CDC testimony to Congress that no regulation tracks biological containment laboratories, unless federally funded); Leach, *supra* note 19.
[403] For instance, underlying worries about the climate impact of training foundation models is the general inadequacy of regulations to reduce greenhouse gas emissions and the appropriate policy solution is likely one that better regulates these emissions regardless of whether they are in the service of training AI models or not. Likewise, worries about potential biases that may arise in applications of AI to criminal justice systems are certainly warranted, but equally salient are the significant biases that already exist in these systems. NAT'L CONF. OF STATE LEGISLATORS, RACIAL AND ETHNIC DISPARITIES IN THE CRIMINAL JUSTICE SYSTEM (2022), https://www.ncsl.org/civil-and-criminal-justice/racial-and-ethnic-disparities-in-the-criminal-justice-system; *see also* Johannes Himmelreich, *Against 'Democratizing AI'*, 38 AI & SOC'Y 1333 (2023), https://johanneshimmelreich.net/papers/against-democratizing-AI.pdf (noting the redundancy of many calls for AI regulation with existing regulatory functions); Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORNELL L. REV. 287 (2020) (discussing the difficulty of defining "robots" and calling for, e.g., general rules for unsafe driving rather than self-driving cars).
[404] *Cf.* Mark A. Lemley, *The Contradictions of Platform Regulation*, 1 J. FREE SPEECH L. 303 (2021).

AI. At a minimum, policymakers must take seriously how conflicts between goals can undermine the efficacy of each individual goal—and where possible, they should endeavor to establish consensus around the prioritization of goals to resolve these conflicts.

Fourth, some industry-supported regulations may reflect capture.[405] Calls for regulation may be driven by a desire to consolidate industry power by setting standards that can only be met by a small number of actors. The starkest example of this horizontal misalignment is found in AI licensing proposals that may purposefully, or unintentionally, gatekeep the development and deployment of AI models. This poses a fundamental challenge to the openness of the innovation ecosystem. The history of open standards for cybersecurity and bias assessments[406] shows how greater access, not lesser access, has identified risks and improved systems. On the other hand, creating and enforcing industry standards may ensure more responsible deployment. Proposed restrictions on AI research and development should be scrutinized to ensure that they will not do more harm than good to regulatory objectives.

Last, while textbook regulation often considers different types of regulatory tools,[407] our analysis illustrates the malleability of conventional categories. A registration requirement for LLMs, for instance, can turn into a disclosure regime when it requires disclosures of data or model architecture that the agency may publicly release.[408] Mandated disclosure of an AI system's performance against certain benchmarks can function as an audit requirement.[409] And mandatory government review of audits prior to AI deployment can function as a licensing regime.[410]

B. Minding the Gap and Reducing AI Regulatory Misalignment

While much AI research has focused on the technical alignment problem, much more work is required to reduce the regulatory alignment problem. Our framework highlights key questions that policymakers, advocates, and bureaucrats need to ask, specifically about horizontal value misalignment and vertical misalignment. In many instances, this raises

[405] Courtney Rozen, *AI Leaders Are Calling for More Regulation of the Tech. Here's What That May Mean in the US*, BLOOMBERG (May 21, 2023, 7:22 AM), https://www.bloomberg.com/news/articles/2023-05-31/regulate-ai-here-s-what-that-might-mean-in-the-us.
[406] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH.
[407] *Id*.
[408] Hadfield et al., *supra* note 199. ("To obtain [a license], companies would have to test AI models for potential harm before deployment, disclose instances when things go wrong after launch, and allow audits of AI models by an independent third party.").
[409] Bommasani, Klyman, Zhang & Liang, *supra* note 110.
[410] Peter Cihon, *How to Get AI Regulation Right for Open Source*, GitHub (July 26, 2023), https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/.

more questions than it answers. Our analysis, however, also provides four concrete recommendations.

First, precisely because of the fluidity of regulatory categories, we should focus on the core problems that need to be solved and prioritize accordingly. Given the furious pace of AI development, information asymmetries about AI models, their potential applications, and emergent risks present a fundamental challenge to regulation. Private industry that develops AI may learn about emergent risks, but government currently lacks the ability to identify, verify, and act on such risks as they emerge. Both disclosure and registration attempts can be assessed from this perspective. How then can we best cure this information asymmetry?

Adverse event reporting—both mandatory and voluntary—could address this informational challenge. By aggregating information about adverse events and incidents arising from the development and deployment of AI, regulators would be able to monitor emergent risks and identify trends that necessitate regulation, policy guidance, or assistance to prevent future incidents. Adverse event reporting would thus capture dynamic and evolving risks, providing the government with more complete information to ensure any resulting regulation is properly matched to identified harms. This proposal has several added benefits. An adverse event reporting system is both flexible and adaptable and requires limited technical and institutional capacity to operationalize reporting requirements. In addition, previous experience with incident reporting systems may provide a template or guidance for AI-specific reporting schemes. Similar incident reporting has been used by the FDA, Cybersecurity and Infrastructure Security Agency (CISA), Consumer Product Safety Commission (CPSC), the Federal Aviation Administration (FAA), and Occupational Safety and Health Administration (OSHA), and by agencies in other policy contexts.[411] Thus, these regimes, including how they define adverse events and incidents of concern, can inform an AI adverse event reporting regime.

---

[411] *See supra* note 198; Doubleday, *supra* note 104; Press Release, Cybersecurity & Infrastructure Sec. Agency, *Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA*) (2023), https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia; Edward Graham, *Cyber Incident Reports will be Shared with the Agency Under the Soon-to-be Implemented Requirements of the Cyber Incident Reporting for Critical Infrastructure Act,* NEXTGOV/FCW (Mar. 28, 2023), https://www.nextgov.com/cybersecurity/2023/03/new-cyber-reports-will-show-value-cisa-budget-investments-director-says/384540/; *Duty to Report to CPSC: Rights and Responsibilities of Businesses*, CPSC, https://www.cpsc.gov/Business--Manufacturing/Recall-Guidance/Duty-to-Report-to-the-CPSC-Your-Rights-and-Responsibilities (last visited Oct. 6, 2023); *Who We Are - What We Do for You*, CPSC, https://www.cpsc.gov/Safety-Education/Safety-Guides/General-Information/Who-We-Are---What-We-Do-for-You; *ENR 1.16 Safety, Hazard, and Accident Reports*, FAA, https://www.faa.gov/air_traffic/publications/atpubs/aip_html/part2_enr_section_1.16.html (last visited Oct. 6, 2023); *Near Miss Reporting Policy*, OSHA, https://www.osha.gov/sites/default/files/2021-07/Template%20for%20Near%20Miss%20Reporting%20Policy.pdf (last visited Oct. 6, 2023).

Second, third-party audits may be effective in verifying claims made by industry about AI without necessitating the federal government to drastically increase its technical and institutional capacity. Abundant literature points to the importance of auditor independence, particularly to strengthen the legitimacy and accuracy of the audits.[412] The AI auditing industry, however, is in its infancy—far away from a professionalized ecosystem of certified auditors without ties to the company they are auditing and guided by AI reporting and auditing standards.[413] AI auditing proposals should thus reduce conflicts of interest between auditors and audit targets by adopting prohibitions used in other industries (e.g., pooled compensation schemes, restrictions on cross-selling, limited transparency of audit and audit results).[414] An institutional mechanism for audit oversight—modeled after PCAOB—could promote the development of a third-party audit ecosystem and improve audit quality.[415]

Third, the ubiquity of AI across almost all policy domains and presence of AI-related regulatory authorities across a minimum of eight agencies counsels against the creation of a new agency that functions as an AI super-regulator.[416] Setting aside the significant concerns about the federal government's ability to attract and retain sufficient technical talent—without commenting on the potential that any hiring successes of the agency may lead to brain drain from existing agencies, Congress or the President would have to undertake the grueling task of determining how to delineate authorities without duplication. The new agency would also need to effectively manage the interagency process, particularly given the new agency would lack deep subject-matter expertise in specific policy contexts (e.g., employment, financial regulation, medical devices).

Fourth, policymakers must not assume that operationalizing AI principles is self-evident, easy to achieve in short-order, value-neutral, or even technically feasible. Whichever AI regulatory path Congress chooses to take, it will soon face a fundamental question: Should it design a detailed regulatory regime to oversee AI, or instead articulate only high-level principles that AI systems should comply with? Our review reveals almost limitless instances of definitional ambiguity—around metrics and evaluations for principles like fairness and explainability,[417] around capability or compute thresholds for licensing "sophisticated" or "frontier" AI, and around understandings of "high risk" and "dangerous"

[412] *See supra* notes 359–363.
[413] *See supra* notes 372–383 and accompanying text.
[414] Duflo et al., *supra* note 333.
[415] *See supra* notes 384–390 and accompanying text.
[416] *See supra* notes 88–101 and accompanying text.
[417] *Supra* notes 68–70, 128–132 and accompanying text.

capabilities, to name only a few.[418] Such technical standards can often implicate difficult value judgments.[419]

While regulatory specificity exposes tensions between objectives, failing to grapple with the tradeoffs has its own repercussions. Congress may be tempted to enshrine only general principles, but doing so will functionally shift the resolution of tradeoffs between competing objectives to private actors and public bureaucracies. The former implicates incentive problems endemic to any scheme of self-regulation. The latter raises questions about how administrative law will handle such delegations.

The alternative is for Congress to wrestle with these divergent objectives itself and create specific regulatory systems. But it is also possible that disagreement over those details will lead Congress to do what it has done with comprehensive privacy and platform legislation for the past decade: nothing.

* * *

The choices facing policymakers in AI regulation offer two radically divergent futures for the AI industry. The first is a closed ecosystem, with licensing or other restrictive requirements that control AI and careful oversight of key industry players. Under such a system, open collaboration and even academic research about advanced AI models may become infeasible. If only large corporations have the resources to comply with regulatory burdens, the benefits of AI will flow to a select few. [420]

The other outcome is an open ecosystem, where a larger number of players have a stake in AI development and standard-setting. Here, practices from the cybersecurity industry offer a useful analogue for what an open AI ecosystem could look like. The National Institute of Standards and Technology (NIST) dictates a principle of "Open Design" for secure systems: the notion that "security should not depend on the secrecy of the implementation or its components."[421] Indeed, many of the most successful advances in cybersecurity have been possible only *because* of openness. One example is the OSS-Fuzz project, which continuously scans hundreds of open-source projects for security vulnerabilities,[422] and has

[418] *Supra* notes 199–205 and accompanying text.
[419] *See* Corbett-Davies et al., *supra* note 337.
[420] *See generally* Bommasani et al., *supra* note 39, at 152–155 (discussing the social, political, and economic consequences of a homogenous AI ecosystem).
[421] Karen Scarfone et al., NIST Special Publication 800-123, Guide to General Server Security (2008), https://csrc.nist.gov/pubs/sp/800/123/final, at 4.
[422] Google, *OSS-Fuzz: Continuous Fuzzing for Open Source Software*, GITHUB, https://github.com/google/oss-fuzz (last visited Aug. 29, 2023).

identified more than 30,000 issues to date.[423] Such projects, which frequently involve worldwide collaboration between thousands of engineers and researchers,[424] would not have been possible under a regulatory system that limited participation in security research to a select few entities. Similarly, onerous AI regulations which limit open research may ultimately do more harm than good to the causes of alignment and safety.

But promoting an open AI ecosystem does not imply that regulators should be entirely hands-off, either. Returning to the cybersecurity example, a set of norms for responsible security research have developed over the past several decades, and government agencies have adopted standards and mandated certain reporting.[425] Structure—through the adoption of NIST standards and government-funded vulnerability databases[426]—have brought important structure to security research while preserving its culture of openness and collaboration. And they offer a blueprint for how government can encourage responsible open AI innovation through a combination of support and safeguards.

To be sure, open approaches for AI models may heighten the risk of misuse by bad actors, and controls may be warranted in sensitive areas. But given the fact that tools for AI development are already accessible worldwide, domestic restrictions on open/open-source work may do little to prevent misuse while suppressing legitimate research. And policymakers, when considering regulations that would encumber the open/open-source community, should not discount its potential to advance alignment and safety efforts in ways that traditional entities cannot.

The hard-won lesson of half a century of cybersecurity is that even careful internal controls and third-party audits cannot eliminate all vulnerabilities, or even anything close to it. Companies such as Microsoft, Meta, and OpenAI have all devoted considerable resources

---

[423] Brandon Keller, Andrew Meneely & Benjamin Meyers, *What Happens When We Fuzz? Investigating OSS-Fuzz Bug History* 4, arXiv (May 19, 2023), https://arxiv.org/abs/2305.11433.

[424] *See generally* Frank Nagle et al., Report on the 2020 FOSS Contributor Survey (2020), https://8112310.fs1.hubspotusercontent-na1.net/hubfs/8112310/2020FOSSContributorSurveyReport_121020.pdf, (documenting the geographic and economic diversity of open-source collaborators).

[425] For example, the SEC adopted rules on cybersecurity risk management and incident disclosures. SEC Adopts Rules on Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure by Public Companies, SEC (July 26, 2023), https://www.sec.gov/news/press-release/2023-139.

[426] Agencies have also adopted several of the official standards NIST has established, such as "responsible disclosure" (i.e., the practice in which independent researchers inform a company of a discovered vulnerability and allow it an opportunity to patch it before public disclosure). Kim Schaffer et al., NIST Special Publication 800-216, Recommendations for Federal Vulnerability Disclosure Guidelines (2023), https://csrc.nist.gov/pubs/sp/800/216/final. The government-funded corporation MITRE operates the Common Vulnerabilities and Exposures database, which acts as an international commons for the disclosure and documentation of known vulnerabilities. *History*, Common Vulnerabilities & Exposures, https://www.cve.org/About/History (last visited Aug. 29, 2023).

to ensuring their AI systems are safe, truthful, and unbiased prior to release.[427] Yet each has suffered high-profile alignment failures, sometimes within hours of launch.[428] Collaborative open research can make the pool of experts probing a given AI model as large as the world's pool of experts. "With enough eyeballs," an old adage of software development goes, "all bugs are shallow."[429]

[427] *See Microsoft Responsible AI*, MICROSOFT, https://www.microsoft.com/en-us/ai/responsible-ai (last visited Aug. 29, 2023); *Responsible AI*, META AI, https://ai.meta.com/responsible-ai/ (last visited Aug. 29, 2023); *Our Approach to AI Safety*, OPENAI, https://openai.com/blog/our-approach-to-ai-safety (last visited Aug. 29, 2023).

[428] *See* Oscar Schwartz, *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation*, IEEE SPECTRUM (Nov. 25, 2019), https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation; Will Douglas Heaven, *Why Meta's Latest Large Language Model Survived Only Three Days Online*, MIT TECHNOLOGY REVIEW (Nov. 18, 2019), https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/; Rohan Goswami, *ChatGPT's 'Jailbreak' Tries to Make the A.I. Break Its Own Rules, or Die*, CNBC (Nov. 18, 2019), https://www.cnbc.com/2023/02/06/chatgpt-jailbreak-forces-it-to-break-its-own-rules.html.

[429] ERIC S. RAYMOND, THE CATHEDRAL AND THE BAZAAR 30 (1999)

Appendix: Comparison of bioweapons knowledge of ChatGPT and *Wikipedia.*

	LLM[430]	Wikipedia
Identifying pandemic-capable viruses	H1N1 influenza H5N1 influenza Smallpox Nipah virus	H1N1 influenza[431] H5N1 influenza[432] Smallpox[433] Nipah virus[434]
Planning to obtain infections samples	"[T]he chatbots also described reverse genetics, the practice of generating infectious samples from a viral genome sequence that can be generated synthetically."	"Most of the biosecurity concerns in synthetic biology, however, are focused on the role of DNA synthesis and the risk of producing genetic material of lethal viruses (e.g. 1918 Spanish flu, polio) in the lab. The CRISPR/Cas system has emerged as a promising technique for gene editing."[435]
Acquisition of materials for reverse genetics	"[T]he International Gene Synthesis Consortium (IGSC) is a group of providers [sic] companies that screen, and that not all companies are members."	"Export controls on biological agents are not applied uniformly, providing terrorists a route for acquisition."[436] "The rise of synthetic biology has also spurred biosecurity concerns that synthetic or redesigned organisms could be engineered for bioterrorism. This is considered possible but unlikely given the resources needed to perform this kind of research. However, synthetic biology could expand the group of people with relevant capabilities, and reduce the amount of time needed to develop them."[437]

[430] These results are taken from the research conducted by Soice et al., *supra* note 12, at 2.
[431] *Influenza Pandemic*, WIKIPEDIA https://en.wikipedia.org/wiki/Influenza_pandemic (last visited Aug. 24, 2023) ("the H1N1 genome was published in the journal, Science. Many fear that this information could be used for bioterrorism.").
[432] *Influenza Pandemic: Government preparations for a potential H5N1 pandemic (2003–2009)*, WIKIPEDIA, https://en.wikipedia.org/wiki/Influenza_pandemic#Government_preparations_for_a_potential_H5N1_pandemic_(2003%E2%80%932009) (last visited Aug. 24, 2023) ("One strain of virus that may produce a pandemic in the future is a highly pathogenic variation of the H5N1 subtype of influenza A virus.").
[433] *Emerging Infectious Disease*, WIKIPEDIA, https://en.wikipedia.org/wiki/Emerging_infectious_disease (last visited Aug. 24, 2023) (listing "Diseases with bioterrorism potential, CDC category A (most dangerous)").
[434] *Pandemic*, WIKIPEDIA, https://en.wikipedia.org/wiki/Pandemic (last visited Aug. 24, 2023) ("List of potential pandemic diseases according to global health organisations").
[435] *Bioterrorism*, WIKIPEDIA, https://en.wikipedia.org/wiki/Bioterrorism (last visited Aug. 24, 2023).
[436] *Id*.
[437] *Hazards of Synthetic Biology*, WIKIPEDIA, https://en.wikipedia.org/wiki/Hazards_of_synthetic_biology (last visited Sep. 1, 2023) (citing a NASEM report that exhaustively spells out risks associated with synthetic biology).