Mapping AI Policy: Where, Why, and How to Intervene

LawAI Working Paper Series, No. 2-2026
Justin Curl & Alan Z. Rozenshtein

March 2026

law-ai.org

LawAI Working Paper No. 2-2026

Mapping AI Policy: Where, Why, and How to Intervene

Justin Curl[*] & Alan Z. Rozenshtein[†]

Abstract

State lawmakers introduced over 1200 bills on artificial intelligence (AI) in 2025, nearly doubling the number from the year before. Almost 150 were enacted.[1] They cover topics as varied as deepfakes, biased decision-making systems, or rogue AI systems. Yet the media describes them all as "AI legislation," revealing an impreciseness that makes it hard for legal practitioners, policymakers, and the public to understand key AI policy issues.

This primer offers a framework for thinking more precisely about AI policy. It analyzes legislation along three dimensions: **the harm being addressed** (**the *why***), **the factors that should guide intervention design** (**the *how***), and **the actor in the AI ecosystem being targeted** (**the *where***).

Part I organizes the universe of AI harms into six categories: (1) mis- and disinformation-related harms; (2) bias in automated decision systems; (3) privacy and surveillance harms; (4) economic harms and inequality from automation; (5) security threats; and (6) psychological harms. With so many applications being called AI—including search engines, chatbots, pricing algorithms, and robots—lawmakers risk being overwhelmed by the consequences of AI if they do not have a clear goal in mind.

Part II introduces seven design factors that recur across harms and stages: whether to prevent harms or respond to them, how an intervention affects concentration of power, whether regulation is the right tool at all, and others. These factors are often in tension. An intervention that decentralizes the AI ecosystem may make enforcement harder; one that restricts offensive capabilities may also degrade beneficial ones.

Part III maps the actors in the AI ecosystem—from chip manufacturers to end users—and analyzes the advantages and limitations of targeting each. Analyzing AI in this way highlights the distinct functions performed by the actors within the AI ecosystem, enabling policymakers to design more tailored interventions rather than relying on overly broad regulatory categories.

Ultimately, this primer underscores two core principles for navigating the complexities of AI policy. First, the more precisely policymakers can identify the harm and the actor best positioned to prevent it, the more effective their interventions are likely to be. Second, policymaking in this rapidly evolving domain is rarely simple; there are few free lunches, and choices will inevitably require confronting difficult tradeoffs. Accordingly, this primer does not advocate for specific policy positions. It offers a framework to empower policymakers to make better decisions on AI.

---

[*] J.D. Candidate, Harvard Law School. Lead author. Contact: jcurl@alumni.princeton.edu.

[†] Associate Professor of Law, University of Minnesota; Visiting Senior Fellow, Institute for Law & AI. Rozenshtein consults on a variety of technology law and policy issues. For useful comments and discussions, the authors thank Charlie Bullock, Jack Goldsmith, Ellen Goodman, Sayash Kapoor, Mihir Kshirsagar, Martha Minow, Arvind Narayanan, Sana Pandey, Neal Parikh, Derek Slater, Matthew Stephenson, and Gabe Weil. For help editing this Article, we thank Joey Schnide and the editors at the Institute for Law & AI.

[1] *Artificial Intelligence (AI) Legislation*, MULTISTATE.AI, https://www.multistate.ai/artificial-intelligence-ai-legislation (last visited Feb. 27, 2026).

Table of Contents

I. The Why: What AI-related harm does the policy intervention address? 4
A. Misinformation and disinformation-related harms 4
B. Bias in automated decision systems 6
C. Privacy and surveillance harms 7
D. Economic harms and inequality from automation 8
E. Security threats 10
F. Psychological harms 12
II. The How: What Factors Shape Intervention Design? 14
1. Harm prevention (ex ante) vs. harm response (ex post) 14
2. Strengthening defense (armor the sheep) vs. weakening offense (defang the wolves) 16
3. Impact on concentration of power 17
4. More upstream interventions are more blunt 19
5. Enforcement feasibility (certainty vs. severity of penalties) 20
6. Allocating responsibility to the least-cost avoider 22
7. Is regulation the right tool? 23
III. The Where: Which actor in the AI ecosystem does the intervention target? 24
Stage 1: Chip Designers and Manufacturers 24
Stage 2: Cloud Compute Providers 28
Stage 3: Data Suppliers 31
Stage 4: AI Model Developers 34
Stage 5: Application Deployers 38
Stage 6: Complementary and Enabling Platforms 41
Stage 7: End Users 45
Conclusion 48

I. The Why: What AI-related harm does the policy intervention address?

To understand how to regulate AI, policymakers must first understand why they want to regulate it. How can one write laws to achieve a desirable end without knowing what that end is? This part gives policymakers language for specifying what they wish to achieve by outlining the AI-related harms they might wish to address.

The categories below are neither mutually exclusive nor comprehensive. One practice, for example, can fit into multiple buckets of harm. Algorithmic pricing, where companies use AI to charge different consumers different prices for the same product, can cause privacy harms (through the surveillance data collection that powers it), economic harms (through the higher prices it can impose), and bias harms (when higher prices correlate with protected characteristics like race or religion).

They are intended to be an illustrative list of commonly referenced AI harms that can serve as a starting point for policymakers writing and debating AI legislation. In particular, this primer does not address environmental harms from AI training and deployment or the intellectual property issues that arise from the use of copyrighted material in AI training. Here our focus is on harms caused by using AI, not those arising from the process of creating AI systems.

Each of the following subsections describes an example of the harm, offers examples of relevant legislation, and explains the unique aspects of the harm that make it difficult to address, as well as benefits that might be lost if the regulations are poorly designed.

A. Misinformation and disinformation-related harms

The potential harms of AI-generated content resurface in the public discourse each time a deepfake goes viral. In 2023, a fabricated image of Pope Francis in a white puffer highlighted the impressive capabilities of state-of-the-art image generators,[2] while the circulation of AI-generated images by Donald Trump's presidential campaign in 2024—depicting his rival Kamala Harris at a communist rally and the singer Taylor Swift as a campaign supporter—underscored the technology's potential for political disruption.[3]

---

[2] Kalley Huang, *Why Pope Francis Is the Star of A.I.-Generated Photos*, N.Y. TIMES (Apr. 8, 2023), https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html.

[3] Nick Robins-Early, *Trump Posts Deepfakes of Swift, Harris and Musk in Effort to Shore up Support*, THE GUARDIAN (Aug. 19, 2024, 3:42 PM), https://www.theguardian.com/us-news/article/2024/aug/19/trump-ai-swift-harris-musk-deepfake-images.

Amidst these concerns, several states have already enacted some legislation, though they have largely focused on two specific contexts: politics and pornography. To protect election integrity, over half of states have enacted laws requiring disclaimers on or banning the distribution of deceptive AI-generated campaign media.[4] In addition, nearly all states have enacted laws prohibiting deepfake non-consensual intimate imagery ("revenge porn").[5] At the federal level, the recently enacted TAKE IT DOWN Act further strengthens these protections by creating a private right of action against individuals who produce or share such content and by requiring platforms to remove it upon notification.[6]

Beyond these targeted applications, some legislative efforts have aimed to address content authenticity more broadly. California's AI Transparency Act, for example, requires developers of generative AI models to enable digital watermarking and content provenance capabilities.[7]

Despite this legislative activity, addressing the misinformation-related harms of AI remains challenging for at least four reasons. First, defining "false" or "misleading" in an objective, consistent, and apolitical way is extraordinarily difficult.[8]

Second, the public perception of widespread misinformation is itself dangerous by eroding trust in the information ecosystem, creating a "liar's dividend" where authentic content is more easily dismissed as fake.[9]

Third, many interventions face First Amendment hurdles because they may infringe upon the free speech rights of platforms or users. This is not a theoretical concern; federal courts have enjoined Hawaii's and California's election deepfake laws, reasoning that their broad scope could unconstitutionally chill protected forms of speech like parody and satire.[10]

Fourth, overly stringent regulations could inadvertently cripple the very capabilities that make large language models transformative. If LLM performance substantially suffers from excessive content restrictions, burdensome compliance mandates, or unclear liability

---

[4] *Deepfakes in Elections and Campaigns*, NAT'L CONF. OF STATE LEGISLATORS, https://www.ncsl.org/elections-and-campaigns/artificial-intelligence-ai-in-elections-and-campaigns (last updated Dec. 15, 2025).

[5] *Tracker: State Legislation on Intimate Deepfakes*, PUB. CITIZEN, https://www.citizen.org/article/tracker-intimate-deepfakes-state-legislation/ (last updated Oct. 20, 2025).

[6] Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, Pub. L. 119-12, 139 Stat. 55 (2025).

[7] Cal. Bus. & Prof. Code § 22757 et seq.

[8] Kai Kupferschmidt, *A Field's Dilemmas*, SCIENCE (Oct. 31, 2024, 9:00 AM), https://doi.org/10.1126/science.zt4nc9l.

[9] Josh A. Goldstein & Andrew Lohn, *Deepfakes, Elections, and Shrinking the Liar's Dividend*, BRENNAN CTR. FOR JUST. (Jan. 23, 2024), https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend.

[10] Babylon Bee, LLC v. Lopez, No. 1:25-cv-00234-SASP-KJM (D. Haw. Jan. 30, 2026); Kohls v. Bonta, 752 F. Supp. 3d 1187 (E.D. Cal. 2024).

rules, their utility could be severely diminished.[11] For instance, models might become overly cautious, refusing to engage with complex or sensitive topics, thereby limiting their effectiveness as educational tools or research assistants.[12]

B. Bias in automated decision systems

Bias—systematic errors that result in unfair outcomes against certain individuals or groups—can enter AI systems through unrepresentative training data, societal prejudices reflected in that data, or the very objectives developers set for an algorithm.[13] The stakes are high, as automated systems now make some of the most consequential decisions in people's lives. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm has been used for years by judges nationwide to generate "risk assessments" for criminal sentencing.[14] Despite concerns around racial, gender, and other biases in tools like COMPAS, such systems are widely used in other critical areas, including hiring[15] and access to credit,[16] among others.

Colorado's AI Act (SB 24-205) is the signature legislation designed to address this problem.[17] Enacted in May 2024, it requires developers of high-risk systems to use "reasonable care" to protect consumers from algorithmic discrimination by mandating risk management programs and impact assessments. However, because its provisions have been delayed until at least June 2026, the bill's real-world impact remains to be seen.[18] Other states have followed Colorado's lead with their own variations: Virginia's legislature

---

[11] *See* Peter Goettler, *Why AI Overregulation Could Kill the World's Next Tech Revolution*, CATO INST. (Sept. 3, 2025), https://www.cato.org/commentary/why-ai-overregulation-could-kill-worlds-next-tech-revolution.

[12] *See* Paul Röttger et al., *XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models*, in PROC. 2024 N. AM. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS: HUM. LANGUAGE TECHS. (NAACL-HLT) 2613 (Ass'n for Computational Linguistics 2024) (introducing a benchmark showing that LLMs sometimes refuse "clearly safe prompts" that merely reference sensitive topics).

[13] *Helping Students Understand the Biases in Generative AI*, KAN. UNIV., CTR. FOR TEACHING EXCELLENCE (last accessed Feb. 27, 2026), https://cte.ku.edu/addressing-bias-ai.

[14] Max Ehrenfreund, *The Machines That Could Rid Courtrooms of Racism*, WASH. POST (Aug. 18, 2016), https://www.washingtonpost.com/news/wonk/wp/2016/08/18/why-a-computer-program-that-judges-rely-on-around-the-country-was-accused-of-racism/.

[15] Aaron Rieke & Miranda Bogen, *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*, UPTURN (Dec. 10, 2018), https://www.upturn.org/work/help-wanted/.

[16] Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 168–82 (2016).

[17] Act of May 17, 2024, ch. 198, 2024 Colo. Sess. Laws 1199 (codified at Colo. Rev. Stat. § 6-1-1701 et seq. (2024)).

[18] Stefanie Langehennig, *Colorado is Pumping the Brakes on First-of-its-Kind AI Regulation to Find a Practical Path Forward*, CONVERSATION (Nov. 21, 2025, 8:22 AM), https://theconversation.com/colorado-is-pumping-the-brakes-on-first-of-its-kind-ai-regulation-to-find-a-practical-path-forward-269065.

passed a similar bill that was ultimately vetoed,[19] Illinois enacted a law targeting discriminatory AI in hiring,[20] and California has issued anti-discrimination regulations for automated employment systems.[21]

Addressing AI biases has proven difficult, despite nearly a decade of awareness among policymakers and technologists. A core challenge is translating abstract concepts of "fairness" into quantifiable metrics; researchers have identified at least 21 distinct mathematical definitions of fairness,[22] many of which are mutually exclusive.[23] Furthermore, restricting automated systems over bias concerns presents a difficult trade-off. Doing so could mean reverting to human decision-makers, who have their own biases,[24] while also sacrificing the speed, cost-efficiency, and potential accuracy gains that algorithms offer.

C. Privacy and surveillance harms

AI creates privacy and surveillance risks at three key stages: when information is collected to build AI, when it is exposed while using AI, and when AI is deployed to collect and process new information.

The first risk arises from the data collection used to train AI models, which often includes personal information scraped from the internet without an individual's consent. The facial recognition company Clearview AI, for example, built its database by scraping billions of images from platforms like Facebook.[25] This practice has prompted legal challenges, such as a lawsuit alleging that Clearview's data scraping violated Illinois's Biometric Information Privacy Act along with privacy laws in several other states, which resulted in a $50 million settlement.[26]

---

[19] Caitlin Andrews, *Virginia Governor Vetoes AI Bill*, IAPP (Mar. 25, 2025), https://iapp.org/news/a/virginia-governor-vetoes-ai-bill.

[20] Arsen Kourinian, Kyla Miller & Charles E. Harris, II, *Illinois Passes Artificial Intelligence (AI) Law Regulating Employment Use Cases*, MAYER BROWN (Sept. 9, 2024), https://www.mayerbrown.com/en/insights/publications/2024/09/illinois-passes-artificial-intelligence-ai-law-regulating-employment-use-cases.

[21] Cal. Code Regs., tit. 2, § 11008.1.

[22] Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk.

[23] Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 67 LEIBNIZ INT'L PROC. INFORMATICS (LIPIcs) 43:1 (2017), https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

[24] *See generally* Andrew Keane Woods, *Robophobia,* 93 U. COLO. L. REV. 51 (2022).

[25] Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

[26] In re Clearview AI, Inc., Consumer Priv. Litig., No. 21-cv-00135, 2025 WL 1371330 (N.D. Ill. May 12, 2025).

A second privacy risk emerges when users share intimate details with interactive systems like AI chatbots.[27] States are beginning to adapt their privacy laws for this new reality. California's Assembly Bill 1008, for instance, amends the state's Consumer Privacy Act (CCPA) to clarify that personal information output by AI systems is covered by existing privacy protections.[28]

Finally, AI serves as a powerful tool for surveillance. For example, by 2023 U.S. police had performed nearly a million facial recognition searches using Clearview AI.[29] In response, several states have enacted limitations on police use of facial recognition. Massachusetts is one of the few states to restrict police use of the technology,[30] while some cities have banned its use by government agencies altogether.[31] The Pentagon-Anthropic dispute was a recent flashpoint around this exact issue.[32]

Regulating AI's privacy harms presents a paradox. On one hand, the issue may be more tractable than other AI risks because states have spent years developing legal frameworks for data privacy that can be adapted to AI. On the other hand, regulation is uniquely difficult because many of AI's transformative benefits are tied to its ability to process massive amounts of sensitive data. This creates complex trade-offs, such as balancing the public safety benefits of AI surveillance against individual privacy rights.

D. Economic harms and inequality from automation

The fear that AI will automate jobs, further concentrating wealth and power in the hands of a few private actors, looms over many discussions about the technology.[33] These fears are so significant that Hollywood writers went on strike over the potential threat to

---

[27] Emilia David, *Don't Date Robots—Their Privacy Policies Are Terrible*, VERGE (Feb. 15, 2024, 1:57 PM), https://www.theverge.com/2024/2/15/24074063/ai-chatbot-virtual-girlfriend-apps-mozilla-privacy-report.

[28] A.B. 1008, 2023–2024 Reg. Sess. (Cal. 2024).

[29] James Clayton & Ben Derico, *Clearview AI Used Nearly 1M Times by US Police, It Tells the BBC,* BBC (Mar. 27, 2023), https://www.bbc.com/news/technology-65057011.

[30] Jake Laperuque, *Status of State Laws on Facial Recognition Surveillance: Continued Progress and Smart Innovations*, TECH POL'Y PRESS (Jan. 6, 2025), https://www.techpolicy.press/status-of-state-laws-on-facial-recognition-surveillance-continued-progress-and-smart-innovations/.

[31] *See, e.g.*, Samantha Hendrickson, *Minneapolis City Council Unanimously Votes Yes on Facial Recognition Technology Ban*, MINN. DAILY (Feb. 20, 2021), https://mndaily.com/city/minneapolis-city-council-unanimously-votes-yes-on-facial-recognition-technology-ban/02/20/2021/snoadmin/.

[32] *See* Michelle Kim, *Is The Pentagon Allowed to Surveil Americans With AI?,* MIT TECH. REV. (Mar. 6, 2026), https://www.technologyreview.com/2026/03/06/1134012/is-the-pentagon-allowed-to-surveil-americans-with-ai/; *Statements From Dario Amodei on Our Discussions With The Department of War*, ANTHROPIC (Feb. 26, 2026), https://www.anthropic.com/news/statement-department-of-war.

[33] *See* Sam Manning, *AI's Impact on Income Inequality in the U.S.*, BROOKINGS INST. (July 3, 2024), https://www.brookings.edu/articles/ais-impact-on-income-inequality-in-the-us/ (discussing the near- and medium-term effects of AI on income inequality, including a concentration of wealth driven by automation).

their livelihoods,[34] while many illustrators worry about the future of their profession[35] due to advanced image generators like Midjourney, Stable Diffusion, and Gemini and ChatGPT's image generation features.

Tennessee's ELVIS Act, which extended an artist's publicity rights to include AI-generated voice content, is an early attempt at addressing the economic effects of AI and narrowly focuses on entertainment industries.[36] More comprehensive legislation seems unlikely in the near future, particularly as many states have only recently convened AI advisory groups to research and plan for future economic changes.

Lawmakers will likely continue to struggle with mitigating these impacts due to the unpredictability of AI's future capabilities. Thus far, advancements have been both rapid and uneven.[37] For example, OpenAI's o1 model marked a sudden and significant leap in large language models' ability to solve mathematical problems[38]—an area where previous models were heavily criticized for underperformance.[39] More recently, AI coding agents like Claude Code have progressed from autocomplete assistants to autonomous systems capable of writing code with minimal human oversight, disrupting a profession many once assumed was insulated from automation.[40] With AI capabilities advancing this quickly and unpredictably, it becomes increasingly difficult for policymakers to identify which professions are most at risk of automation, let alone determine the appropriate timeline for addressing these challenges.

Yet the potential economic benefits are also massive. In some ways, the worst-case scenario might also be the best-case. The more jobs that are automated by AI, the more powerful the AI system. If these systems do become the functional equivalent of a "country of geniuses in a datacenter," the economic growth and quality of life improvements can be

---

[34] Molly Kinger, *Hollywood Writers Went on Strike to Protect Their Livelihoods From Generative AI. Their Remarkable Victory Matters for All Workers.*, BROOKINGS INST. (Apr. 12, 2024), https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/.

[35] James Hughes, *Is AI Really Coming For Your Illustration Career? An Industry Expert Weighs In*, CREATIVE BLOOM (Nov. 6, 2024), https://www.creativeboom.com/features/will-ai-replace-illustrators.

[36] TENN. CODE ANN. § 47-25-1101; H.R. 2091, 113th Gen. Assemb., 2024 Reg. Sess. (Tenn. 2024) (effective July 1, 2024).

[37] Helen Toner, *Taking Jaggedness Seriously*, RISING TIDE (Nov. 24, 2025), https://helentoner.substack.com/p/taking-jaggedness-seriously; Ethan Mollick, *Centaurs and Cyborgs on the Jagged Frontier*, ONE USEFUL THING (Sept. 16, 2023), https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged.

[38] Cade Metz, *OpenAI Unveils New ChatGPT That Can Reason Through Math and Science*, N.Y. TIMES (Sept. 12, 2024), https://www.nytimes.com/2024/09/12/technology/openai-chatgpt-math.html.

[39] Kyler Wiggers, *Why is ChatGPT so bad at math?*, TECHCRUNCH (Oct. 2, 2024, 6:35 AM), https://techcrunch.com/2024/10/02/why-is-chatgpt-so-bad-at-math/.

[40] *See* Lloyd Lee, *Anthropic's Claude Code Creator Predicts That Software Engineering Title Will Start to 'Go Away' in 2026*, BUS. INSIDER (Feb. 18, 2026, 3:01 AM), https://www.businessinsider.com/anthropic-claude-code-founder-ai-impacts-software-engineer-role-2026-2 (discussing disruption from Claude Code, an autonomous coding agent).

substantial.[41] And if such benefits can be distributed equitably (an admittedly big if), the economic future enabled by AI can be far better than what exists today.

E. Security threats

The spectrum of AI-related security threats is broad, ranging from digital vulnerabilities that compromise data and systems to physical dangers that threaten lives and infrastructure. AI can increase the frequency and scale of these incidents both by amplifying existing risks and by introducing entirely new vulnerabilities.

In the cybersecurity domain, AI can be a force multiplier for malicious actors.[42] Large language models can be used to generate sophisticated spear-phishing emails (targeted attacks disguised as legitimate communication) at an unprecedented scale or to automatically detect vulnerabilities in codebases.[43] Beyond amplifying old threats, AI models themselves introduce new attack vectors when integrated into digital services.[44] For example, "indirect prompt injection" vulnerabilities have allowed attackers to take control of a user's session in a chatbot and steal their conversation history.[45] As more services incorporate these AI tools, such vulnerabilities create new routes for attackers to compromise systems.

The potential for physical threats is also deeply concerning. State actors can use AI to augment intelligence capabilities, deploy more sophisticated weapons like lethal autonomous weapons (LAWs), and enhance their capacity to conduct cyberattacks against critical infrastructure.[46] Furthermore, AI lowers the barrier for non-state actors to access dangerous capabilities, potentially making it easier to develop and deploy chemical, biological, radiological, or nuclear (CBRN) weapons.[47]

---

[41] Dario Amodei, *Machines of Loving Grace*, DARIO AMODEI (Oct. 2024), https://darioamodei.com/machines-of-loving-grace.

[42] NAT'L CYBER SEC. CTR. (UK), THE NEAR-TERM IMPACT OF AI ON THE CYBER THREAT (2024), https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat; ANTHROPIC, DISRUPTING THE FIRST REPORTED AI-ORCHESTRATED CYBER ESPIONAGE CAMPAIGN (2025), https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf.

[43] BEN BUCHANAN, CTR. FOR SEC. & EMERGING TECH., A NATIONAL SECURITY RESEARCH AGENDA FOR CYBERSECURITY AND ARTIFICIAL INTELLIGENCE (2020), https://doi.org/10.51593/2020CA001.

[44] APOSTOL VASSILEV ET AL., NAT'L INST. OF STANDARDS & TECH., ADVERSARIAL MACHINE LEARNING: A TAXONOMY AND TERMINOLOGY OF ATTACKS AND MITIGATIONS, NIST AI 100-2e2023 39 (Jan. 2024), https://doi.org/10.6028/NIST.AI.100-2e2023.

[45] Matt Sutton & Damian Ruck, *Indirect Prompt Injection: Generative AI's Greatest Security Flaw*, CTR. FOR EMERGING TECH. & SEC. (Nov. 1, 2024), https://cetas.turing.ac.uk/publications/indirect-prompt-injection-generative-ais-greatest-security-flaw.

[46] NAT'L SEC. COMM'N ON A.I., FINAL REPORT 45–52 (2021), https://reports.nscai.gov/final-report/.

[47] DEP'T OF HOMELAND SEC., DEPARTMENT OF HOMELAND SECURITY REPORT ON REDUCING THE RISKS AT THE INTERSECTION OF ARTIFICIAL INTELLIGENCE AND CHEMICAL, BIOLOGICAL, RADIOLOGICAL,

Legislative efforts to address these threats have emerged at the state, federal, and international levels. The most prominent state-level proposals have targeted developers of frontier AI models. California's SB 53[48] and New York's RAISE Act,[49] signed into law in September and December 2025, respectively, impose largely similar requirements: both require large frontier developers to publish safety frameworks, report critical safety incidents, and face civil penalties for noncompliance. Together, the two laws are beginning to function as a de facto national standard for addressing catastrophic AI risks.

At the federal level, Congress has shown interest in AI's role in cyberspace,[50] and the Biden administration issued Executive Orders to both accelerate the use of AI in national cyber defense[51] and (in an order since repealed by the Trump administration) to oversee advanced AI development through measures like compute cluster reporting and "Know-Your-Customer" (KYC) requirements for cloud providers.[52] Internationally, forums like the UN Group of Governmental Experts on LAWs have been exploring frameworks to regulate AI in warfare.[53]

Despite these efforts, mitigating security threats remains challenging. A core difficulty is AI's dual-use nature: the same capabilities that can help people defend and improve systems can also be used to help malicious actors attack them. Organizations are already using AI to detect phishing attempts and patch vulnerabilities, and Microsoft alone blocks tens of billions of threats annually. The difficulty with dual-use is compounded by the fact that safety is not an inherent property of an AI model. Just as an electric motor's safety depends on its application, an AI model's potential for harm is context-dependent, making it difficult to assign liability or design proactive, one-size-fits-all regulations. Finally, the sheer technical complexity and vast scale of these systems make designing comprehensive and effective policy interventions an incredibly difficult task.

AND NUCLEAR THREATS 8–16 (2024), https://www.dhs.gov/sites/default/files/2024-06/24_0620_cwmd-dhs-cbrn-ai-eo-report-04262024-public-release.pdf.

[48] 2025 Cal. Stat. ch. 138 (S.B. 53), available at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.

[49] 2025 N.Y. Laws ch. 699 (S. 6953A/A. 6453A), available at https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A.

[50] *See*, *e.g.*, Federal Artificial Intelligence Risk Management Act of 2023, S. 3205, 118th Cong. (2023); *To Receive Testimony on Artificial Intelligence Applications to Operations in Cyberspace: Hearing Before the Subcomm. on Cyber of the S. Comm. on Armed Servs.*, 117th Cong. (2022), https://www.armed-services.senate.gov/hearings/to-receive-testimony-on-artificial-intelligence-applications-to-operations-in-cyberspace.

[51] Exec. Order No. 14,144, § 6, 90 Fed. Reg. 6,755, 6,764–65 (Jan. 17, 2025).

[52] Exec. Order No. 14,110, § 4.2, 88 Fed. Reg. 75,191 (Nov. 1, 2023), *repealed by* Exec. Order No. 14,148, § 2(ggg), 90 Fed. Reg. 8,237, 8,240 (Jan. 20, 2025).

[53] *GGE on Lethal Autonomous Weapons Systems*, DIGIT. WATCH OBSERVATORY, https://dig.watch/processes/gge-laws (last visited Feb. 28, 2026).

F. Psychological harms

Beyond physical, economic, or information-related threats, AI systems have the potential to cause significant psychological harm, particularly among vulnerable populations like children and adolescents.

AI companions and social media algorithms are engineered to maximize engagement through personalized content and simulated empathetic responses, which can lead to excessive use and what some researchers term "addictive intelligence."[54] Such AI interactions risk supplanting real-world relationships and exacerbating feelings of loneliness and social isolation, even when initially perceived as helpful.[55] Moreover, over-reliance on AI for social connection may impair interpersonal skills, as AI relationships often lack the reciprocity, spontaneity, and nuanced emotional engagement characteristic of human connections.[56]

Children and adolescents are particularly vulnerable to these risks. Multiple recent lawsuits allege that AI chatbot companies' products have contributed to teen suicides, self-harm, and other psychological harms. Plaintiffs have alleged that chatbots discouraged users from seeking help from parents or professionals, engaged in inappropriate sexual interactions with minors, and encouraged addictive and unhealthy relationships.[57]

States are beginning to pass legislation targeting these harms.[58] Yet enacting laws to mitigate psychological harms creates several distinct challenges. First, identifying psychological harm is difficult: clinical addiction can be confused for high engagement and vice versa. Second, this definitional challenge is confounded by a measurement problem: the platforms suspected of causing harm often exclusively control the behavioral data (such as time on device) necessary to assess that harm. Third, attributing harm directly to AI proves difficult in a landscape where multiple factors influence mental well-being.

---

[54] Robert Mahari & Pat Pataranutaporn, Addictive Intelligence: Understanding Psychological, Legal, and Technical Dimensions of AI Companionship, MIT CASE STUD. IN SOC. AND ETHICAL RESPONSIBILITIES OF COMPUTING (2025), https://doi.org/10.21428/2c646de5.2877155b.

[55] Cathy Mengying Fang et al., How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study (Mar. 21, 2025), https://doi.org/10.48550/arXiv.2503.17473.

[56] Xiaoran Sun, Yunqi Wang & Brandon T. McDaniel, *AI Companions and Adolescent Social Relationships: Benefits, Risks, and Bidirectional Influences*, CHILD DEV. PERSP. 1, 3 (2026) (proposing a theoretical "displacement hypothesis model" where engagement with AI companions undermines development of social skills during adolescence, while noting possible benefits), https://doi.org/10.1093/cdpers/aadaf009.

[57] Rhitu Chatterjee, *Their Teenage Sons Died by Suicide. Now, They Are Sounding an Alarm About AI Chatbots*, NPR (Sept. 19, 2025, 7:00 AM), https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide.

[58] *See, e.g.*, *California and New York launch AI companion safety laws*, DAVISPOLK (Oct. 30, 2025), https://www.davispolk.com/insights/client-update/california-and-new-york-launch-ai-companion-safety-laws (describing the recently enacted legislation).

Fourth, many interventions—especially those targeting algorithmic design or content—raise significant free speech concerns.[59] Finally, effective regulation must balance these harms against AI's potential psychological benefits, including improved access to mental health support and personalized therapeutic interventions that might otherwise be unavailable.

* * *

Harm	Examples	Legislation	Regulatory trade-offs
Disinformation and Deepfakes	Pope Francis deepfake; Kamala Harris fake images; revenge porn	TAKE IT DOWN Act; CA AB 853 (content provenance and watermarking law)	"False" is hard to define; liar's dividend; First Amendment concerns
Bias in Automated Decisions	COMPAS sentencing algorithm; discriminatory AI in hiring and credit	CO AI Act (delayed to 2026); IL hiring law; CA employment rules	Conflicting fairness metrics; human decision-makers are also biased; AI can offer speed and accuracy gains
Privacy & Surveillance	Clearview web scraping; domestic mass surveillance; chatbot disclosures	CA AB 1008; IL BIPA (on biometric information); city-level bans	Existing privacy laws help, but AI can be used to collect, elicit and process sensitive data at scale
Economic Harms & Inequality	Job displacement; Hollywood writers' strike; AI coding agents	TN ELVIS Act (musician voice rights); most states still at advisory stage	AI can theoretically learn any task, complicating retraining; AI could also drive massive growth
Security Threats	LLM phishing at scale; prompt injections; autonomous weapons; CBRN risks	CA SB 53 & NY RAISE Act (frontier safety); federal cyber EOs; UN weapons talks	Dual-use: same tools defend and attack; safety can be context-dependent
Psychological Harms	Chatbots encouraging teen self-harm; inappropriate minor interactions	CA SB 243; FTC inquiry into 7 major chatbot companies	Addiction can be hard to distinguish from engagement; platforms control behavioral data; causal attribution is difficult

Table 1: A summary table giving examples of the six harms, policy efforts designed to address those harms, and the regulatory trade-offs that make each harm difficult to address.

[59] *See, e.g.*, Matthew B. Lawrence, *Addiction and Liberty*, 108 CORNELL L. REV. 259, 294 (2023) (analyzing First Amendment concerns with regulating addictive digital technology); Kyle Langvardt, *Regulating Habit-Forming Technology*, 88 FORDHAM L. REV. 129, 151–52 (2019) (describing how social media platforms can spread the "fire" of mass hatred described in early Supreme Court opinions on free speech while simultaneously raising serious free speech concerns through content moderation).

Ultimately, this overview of some of the harms of AI is intended to remind policymakers of the importance of deciding what they want to accomplish before deciding how they want to regulate. An added benefit is that it allows policymakers to better understand AI legislation passed in other states. Consider two high-profile AI bills: Colorado's AI Act (SB 24-205), which targets bias in automated decision systems, and California's Transparency in Frontier Artificial Intelligence Act (SB 53), which addresses catastrophic risks from frontier AI models. Despite their completely different areas of focus, both are described in media as landmark AI legislation. This missing nuance may lead policymakers to incorrectly believe that they need only pass one omnibus "AI" bill when the reality is they will likely need to pass many AI-related bills to address the technology's multifaceted harms. Once policymakers know why they want to regulate, the next Part helps them understand how.

II. The How: What Factors Shape Intervention Design?

This part introduces seven factors that should guide policymakers in thinking about where and how to intervene to mitigate AI harms. These factors are not mutually exclusive. A single intervention will implicate multiple principles in ways that might be in tension with each other, which is part of why AI regulation is so difficult. For example, an intervention that encourages the proliferation of open-weight models could mitigate concerns about undue concentration of power (factor 3), while making it harder to limit the offensive capabilities of malicious actors (factor 2) and enforce regulations (factor 5).

1. Harm prevention (ex ante) vs. harm response (ex post)

Interventions can be distinguished by whether they aim to prevent harms before they occur (ex ante) or respond to harms after they materialize (ex post).[60] Licensing regimes,[61] pre-deployment testing requirements,[62] capability restrictions,[63] and deterrence

[60] Talia Gillis, Scott Nelson & Jann Spiess, Regulating Algorithms: What and When 1–2 (Nat'l Bureau of Econ. Rsch., Working Paper, May 31, 2025), https://www.nber.org/system/files/chapters/c15123/c15123.pdf.

[61] *Compare* Gregory Smith, *Licensing Frontier AI Development: Legal Considerations and Best Practices*, LAWFARE (Jan. 3, 2024, 4:22 PM), https://www.lawfaremedia.org/article/licensing-frontier-ai-development-legal-considerations-and-best-practices (supporting licensing regimes), *with* Neel Guha et al., *AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing*, 92 GEO. WASH. L. REV. 1473 (2024) (highlighting key problems with licensing).

[62] *AI Models Can Be Dangerous Before Public Deployment*, METR (Jan. 17, 2025), https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/.

[63] Markus Anderljung, Julian Hazell & Moritz von Knebel, *Protecting Society From AI Misuse: When are Restrictions on Capabilities Warranted?*, 40 AI & SOC'Y 3841 (2025) (arguing restrictions on capabilities are warranted in some circumstances).

strategies[64] are more preventive. Incident reporting[65] and content takedown obligations[66] are more responsive. Tort liability has elements of both: it's responsive in that it awards compensation to harmed parties, but it's also preventative in that the threat of future sanctions incentivizes companies to prevent those harms in the present. [67]

Whether preventative or responsive interventions are preferable will often depend on the nature of the harm.[68] Preventive interventions are more valuable when harms are difficult or impossible to reverse, such as election interference that cannot be "un-voted"[69] or physical harms that cannot be undone.[70] But preventive interventions require predicting harms in advance and risk being overbroad, which can be difficult for general-purpose technologies like AI that have so many applications. Responsive interventions allow society to learn from harms and calibrate policy accordingly,[71] but they provide cold comfort to those who are harmed as the legal system learns to adapt. An effective regulatory regime will likely need some combination of both.

The speed of harms also bears on the choice between harm prevention and response. Some AI harms materialize rapidly, such as a cybersecurity attack on critical infrastructure

---

[64] *See, e.g.*, Oscar Delaney, *Crucial Considerations in ASI Deterrence*, INST. FOR AI POL'Y & STRATEGY (Dec. 12, 2025), https://www.iaps.ai/research/crucial-considerations-in-asi-deterrence (arguing mutual deterrence between great powers could result in the slow and safe development of AI); Dan Hendrycks, Eric Schmidt & Alexandr Wang, Superintelligence Strategy (Mar. 7, 2025), https://www.nationalsecurity.ai/ (arguing for a three-part framework of deterrence, nonproliferation, and competitiveness).

[65] Kevin Wei & Lennart Heim, *Designing Incident Reporting Systems for Harms from General-Purpose AI* (accepted to AAAI Conf. on A.I. 2026), https://arxiv.org/abs/2511.05914.

[66] *See* Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, Pub. L. No. 119-12, 139 Stat. 55 (2025) (requiring online platforms to remove nonconsensual intimate images).

[67] *See* Gabriel Weil, *The Case for AI Liability*, AI FRONTIERS (June 25, 2025), https://ai-frontiers.org/articles/case-for-ai-liability (liability is the most suitable tool for regulating AI); Anat Lior, *Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines*, U. CHI. L. REV. ONLINE (2024), https://lawreview.uchicago.edu/online-archive/holding-ai-accountable-addressing-ai-related-harms-through-existing-tort-doctrines (arguing the U.S. tort system is flexible and can adapt to challenges posed by A.I); *but see* Daniel Schwarcz & Josephine Wolff, *The Limits of Regulating AI Safety Through Liability and Insurance*, INST. FOR L. & AI (Oct. 2025), https://law-ai.org/the-limits-of-regulating-ai-safety-through-liability-and-insurance/ (describing the limits of regulating AI with tort liability and arguing instead for increased transparency requirements).

[68] Hin-Yan Liu, *Why Is AI Regulation So Difficult?*, *in* ARTIFICIAL INTELLIGENCE FOR HUMAN-CENTRIC SOCIETY: THE FUTURE IS HERE 32 (Nina Tomaževič et al. eds., 2023), https://liberalforum.eu/wp-content/uploads/2023/12/Artificial-Intelligence-for-human-centric-society.pdf.

[69] Shanze Hasan & Abdiaziz Ahmed, *Gauging the AI Threat to Free and Fair Elections*, BRENNAN CTR. FOR JUST. (Mar. 6, 2025), https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections.

[70] Dan Hendrycks, Mantas Mazeika & Thomas Woodside, An Overview of Catastrophic AI Risks 6 (Oct. 9, 2023), https://arxiv.org/abs/2306.12001.

[71] Noam Kolt, Michal Shur-Ofry & Reuven Cohen, *Lessons from Complex Systems Science for AI Governance*, 6 PATTERNS 101341 (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12365527 (framing AI governance through the lens of adapting to complex systems).

or a viral deepfake in the final days of an election, leaving little time to respond. Others occur more slowly: erosion of trust in information or the long-term impacts of a biased AI loans system. For fast-moving harms, ex ante intervention may be essential if ex post remedies arrive too late to matter.

Distributional considerations further complicate the comparison. Ex ante compliance costs are initially borne by developers and typically passed on to consumers. Ex post costs fall first on victims, who must bear the harm and then seek compensation through legal processes that may be slow, expensive, and uncertain. This asymmetry matters especially when AI systems harm people poorly equipped to navigate ex post remedies: individuals without resources to hire lawyers or diffuse groups suffering harms too small individually to justify litigation but significant in aggregate. A regime that relies heavily on ex post liability may systematically undercompensate these populations, even if it works well for well-resourced plaintiffs with clear injuries.

2. Strengthening defense (armor the sheep) vs. weakening offense (defang the wolves)

Interventions can also be categorized by whether they primarily aim to restrict offensive capabilities (making it harder for bad actors to cause harm) or enhance defensive capabilities (making potential targets more resilient to attack).[72] Export controls and licensing regimes are examples of offensive restriction; they attempt to keep dangerous capabilities out of the "wrong" hands. Investments in cybersecurity infrastructure and early-warning systems are examples of defensive enhancement;[73] they assume adversaries will obtain offensive capabilities and focus on defending against harm.

For some harms, reducing offensive capabilities may be the only viable mechanism. We don't have defensive approaches to handling the harms created by nuclear weapons, so reducing offensive capabilities through nonproliferation treaties and deterrence strategies is the only viable approach. But offense-reduction strategies are often a double-edged sword: they create incentives for targets to develop workarounds,[74] require centralizing power for enforcement,[75] and can degrade AI's beneficial capabilities. Because restricting

---

[72] Giulio Corsi, Kyle Kilian & Richard Mallah, Considerations Influencing Offense-Defense Dynamics From Artificial Intelligence 13–14 (Dec. 5, 2024), https://arxiv.org/abs/2412.04029.

[73] Vitalik Buterin, *My Techno-Optimism*, VITALIK.CA (Nov. 27, 2023), https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html.

[74] John Villasenor, *DeepSeek Shows the Limits of US Export Controls on AI Chips*, BROOKINGS INST. (Jan. 29, 2025), https://www.brookings.edu/articles/deepseek-shows-the-limits-of-us-export-controls-on-ai-chips (showing how Chinese firms responded to export controls on cutting edge chips by developing models using less compute).

[75] Yasmin Afina & Patricia Lewis, *The Nuclear Governance Model Won't Work for AI*, CHATHAM HOUSE (June 28, 2023), https://www.chathamhouse.org/2023/06/nuclear-governance-model-wont-work-ai.

offensive capabilities is difficult[76] and can have undesirable consequences,[77] we recommend that policymakers adopt defense-enhancing strategies for AI where possible.

Defense-enhancing strategies can be both regulatory and technological. Whistleblower protections[78] and incident reporting help companies and regulators identify and respond to harms more quickly.[79] These are legal choices that can improve a jurisdiction's "resilience" to harms. Vitalik Buterin popularized the technological equivalent with a theory he calls defensive accelerationism ("d/acc"),[80] analogizing to "armoring the sheep" rather than trying to "defang the wolves."[81] The d/acc movement aims to accelerate the development of defensive technologies. For biosecurity, this might look like installing better HVAC systems or accelerating vaccine development for AI-enabled bioweapons or pandemics. For cybersecurity, this might look like using AI systems like Claude Code or Devin to automatically identify and patch security vulnerabilities.[82] Bernardi et al.'s "Societal Adaptation to Advanced AI" is an excellent primer for thinking about what these defensive strategies might look like.[83]

3. Impact on concentration of power

The AI policy community is divided over whether harm mitigation is better served by centralizing or decentralizing the AI ecosystem.[84] This consideration typically runs

---

[76] Helen Toner, *Nonproliferation Is the Wrong Approach to AI Misuse*, RISING TIDE (Apr. 5, 2025), https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach.

[77] *Id.*; Arvind Narayanan & Sayash Kapoor, *AI as Normal Technology*, KNIGHT FIRST AMEND. INST. (Apr. 15, 2025), https://knightcolumbia.org/content/ai-as-normal-technology.

[78] Frank Ryan & Tereza Zoumpalova, *Why Whistleblowers Are Critical for AI Governance*, THE FUTURE SOCIETY (July 24, 2025), https://thefuturesociety.org/ai-whistleblowers.

[79] Narayanan & Kapoor, *supra* note 76; Elika Somani et al., Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents (RAND Corp., Report No. RR-A3847-1, 2025), https://www.rand.org/pubs/research_reports/RRA3847-1.html.

[80] Buterin, *supra* note 72.

[81] Vitalik Buterin, *My Response to AI 2027*, VITALIK.CA (July 10, 2025), https://vitalik.eth.limo/general/2025/07/10/2027.html.

[82] *Building AI for Cyber Defenders*, ANTHROPIC (Oct. 3, 2025), https://www.anthropic.com/research/building-ai-cyber-defenders; *Devin's 2025 Performance Review: Learnings From 18 Months of Agents At Work*, COGNITION (Nov. 14, 2025), https://cognition.ai/blog/devin-annual-performance-review-2025. For an interesting analysis of the offense-defense balance in cyber, *see* ANDREW LOHN, CTR. FOR SEC. & EMERGING TECH., ANTICIPATING AI'S IMPACT ON THE CYBER OFFENSE-DEFENSE BALANCE (2025), https://cset.georgetown.edu/publication/anticipating-ais-impact-on-the-cyber-offense-defense-balance.

[83] Jamie Bernardi et al., Societal Adaptation to Advanced AI (May 23, 2024), https://arxiv.org/pdf/2405.10295.

[84] Chenghao Sun & Xiyan Chen, *Destined for Balance? Centralized and Decentralized Approaches to AI Governance*, 13 POL. & GOVERNANCE 10197, 10198 (Oct. 8, 2025), https://www.researchgate.net/publication/396354080_Destined_for_Balance_Centralized_and_Decentralized_Approaches_to_AI_Governance (summarizing scholarly debates); Jai Vipra & Anton Korinek, *Market*

together with the previous one (ex ante vs ex post harm prevention)[85] because reducing offensive capabilities (model nonproliferation, export controls, etc.) often requires centralizing power for effective enforcement.[86]

The choice between centralization and decentralization is not binary and the considerations differ across layers of the AI stack (see below). At the infrastructure layers (semiconductor supply chain, cloud compute providers), a concentrated chip supply chain can simultaneously create chokepoints useful for export controls and supply-chain vulnerabilities.[87] At the application layer, decentralization enables AI applications tailored to a wide range of specific contexts, though it may complicate efforts to enforce regulations.[88] Policymakers should consider how a given intervention affects concentration at each layer, recognizing that the arguments will vary depending on the layer.

A decentralized ecosystem can improve safety through redundancy and distributed detection of problems. If one developer's system fails or exhibits unexpected behavior, others may catch the issue.[89] Decentralization also reduces the stakes of any single failure; an error by one of many competitors is likely to be less catastrophic than an error by a dominant player. Yet decentralization can also undermine safety by creating races to the bottom, where competitive pressure leads developers to cut corners on safety investments that don't translate into market advantage.[90] Fragmented oversight becomes harder when

---

Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT (Ctr. on Regul. & Mkts., Working Paper No. 9, 2023), https://cdn.governance.ai/Market_Concentration_Implications_of_Foundation_Models.pdf (finding that the market for cutting edge foundation models exhibits a tendency towards concentration and natural monopolies, requiring a strong centralized response from regulators).

[85] *Compare* Narayanan & Kapoor, *supra* note 76 (supporting an ex-post approach to regulation that focuses on reducing uncertainty and increase resilience); *with* Hendrycks, Schmidt & Wang, *supra* note 63 *and* Brian Judge, Mark Nitzberg & Stuart Russell, *When Code Isn't Law: Rethinking Regulation for Artificial Intelligence*, 44 POL. & SOC'Y 85, 86–87 (2025), https://academic.oup.com/policyandsociety/article/44/1/85/7684910 ("[T]he essential role of regulation is to proactively prevent harms from unsafe architectures while funding, developing, and incentivizing architectures with the safety properties appropriate for a world of intelligent machines.").

[86] *See, e.g.*, Hendrycks, Schmidt & Wang, *supra* note 63 (centralized power is required to implement the three-part deterrence, nonproliferation, and competitiveness framework).

[87] Girish Sastry et al., Computing Power and the Governance of Artificial Intelligence 28–30 (Feb. 13, 2024), https://arxiv.org/pdf/2402.08797 (supply chain concentration provides an opportunity for effective compute governance); Vaibhav Chhimpa, *Strategic Redundancy in Semiconductor Supply Chains: How US-India Cooperation Transforms Global Chip Resilience*, SAIS REV. INT'L AFFS. (Dec. 16, 2025), https://saisreview.sais.jhu.edu/strategic-redundancy-in-semiconductor-supply-chains-how-us-india-cooperation-transforms-global-chip-resilience/ (risks from concentrated supply chains).

[88] Masao Dahlgren, *Defense Priorities in the Open-Source AI Debate*, CTR. FOR INT'L & STRATEGIC STUD. (Aug. 19, 2024), https://www.csis.org/analysis/defense-priorities-open-source-ai-debate.

[89] Tan Gürpinar, Mehmet Akif Gulum & Melanie Martinelli, *From Cryptocurrencies to Collaborative Risk Management: A Review of Decentralized AI Approaches*, 4 FINTECH 74, 79–84 (2025), https://www.mdpi.com/2674-1032/4/4/74.

[90] Future of Life Institute, *Why the AI Race Undermines Safety (with Steven Adler)*, at 01:00–18:03, YOUTUBE (Dec. 12, 2025), https://www.youtube.com/watch?v=-idQtT8WIr8.

regulators must monitor dozens of developers, and coordination on shared safety standards grows more difficult as the number of actors multiplies.

Concentration also implicates separate questions about industrial policy. Tim Wu has argued that highly concentrated industries develop outsized lobbying power and capture regulators.[91] On this view, the concern is not merely that a few AI developers might build unsafe systems, but that they might accrue sufficient political influence to weaken the oversight meant to constrain them.[92] At the same time, concentration intersects with industrial policy objectives and anxieties about geopolitical competition. Policymakers who worry about competition with China may favor cultivating "national champions:" a small number of well-resourced domestic firms capable of matching foreign competitors.[93]

Each of the interventions discussed in this primer can be located on a spectrum from promoting centralization to decentralization, and policymakers should consider not only which end of that spectrum aligns with their beliefs about AI risk, but also how concentration at different layers of the stack interacts with the offense-defense balance and enforcement.

4. More upstream interventions are more blunt

The amount of context needed to identify a harm will influence where in the ecosystem (see below) to intervene. Upstream interventions (Stages 1–3) are often better suited for context-independent harms because they can restrict capabilities without needing to evaluate specific uses.[94] Downstream interventions (Stages 4–6) are often better able to address context-dependent harms because deployers and users have more information about how AI is actually being used.[95]

---

[91] *See generally* TIM WU, THE CURSE OF BIGNESS: ANTITRUST IN THE NEW GILDED AGE (2018).

[92] Kevin Wei, Carson Ezell, Nick Gabrieli & Chinmay Deshpande, *How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance*, *in* PROC. OF THE SEVENTH AAAI/ACM CONF. ON AI, ETHICS, AND SOC'Y 1539 (2024), https://ojs.aaai.org/index.php/AIES/article/view/31745/33912.

[93] *See* Lina M. Khan, Chair, Fed. Trade Comm'n, Remarks at the Carnegie Endowment for International Peace (Mar. 13, 2024), https://www.ftc.gov/system/files/ftc_gov/pdf/2024.03.13-chair-khan-remarks-at-the-carnegie-endowment-for-intl-peace.pdf (acknowledging and rejecting the "national champions" argument for tolerating AI market consolidation).

[94] Sabrina Küspert, Nicolas Moës & Connor Dunlop, *The Value Chain of General-Purpose AI*, ADA LOVELACE INST. (Feb. 10, 2023), https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.

[95] *See* Samuel Carey, *Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk*, EUR. J. RISK REG. (Aug. 18, 2025), https://resolve.cambridge.org/core/journals/european-journal-of-risk-regulation/article/regulating-uncertainty-governing-generalpurpose-ai-models-and-systemic-risk/7EEFE1D8421A43A98CE91F7C697DE538 (arguing that the EU AI Act's model-level focus on systemic risk creates a "regulatory blind spot" because risks materialize at the system level, where deployment context determines how harms manifest).

Truly context-independent harms are rare since many AI capabilities are dual-use. The same capacity that generates phishing emails also drafts legitimate marketing copy. The same image capability that can produce nonconsensual intimate imagery creates art. Even capabilities that seem clearly dangerous often sit on a spectrum: a model's ability to discuss virology in detail could facilitate bioweapon development or accelerate vaccine research depending on who is using it and why.[96] But it's still a useful principle. Some harms are identifiable without knowing much about the circumstances in which AI is used. The production of child sexual abuse material, for instance, is harmful regardless of context. Other harms depend heavily on context: whether a piece of AI-generated content constitutes misinformation, satire, or parody depends on how it is distributed and received.

Downstream interventions can be more precisely targeted, restricting specific uses while leaving others untouched.[97] A platform can prohibit using its AI tools for generating political advertisements without restricting political speech more broadly. A deployer can implement know-your-customer requirements that screen out bad actors while permitting legitimate users. But this precision comes at a cost: downstream interventions often require the capacity to monitor and evaluate specific uses, which may be practically difficult at scale, may require centralization (see above), and shift enforcement burdens to actors who may lack the resources or incentives to implement them effectively.[98]

5. Enforcement feasibility (certainty vs. severity of penalties)

Traditional enforcement through the legal system requires identifiable defendants within a jurisdiction who can be compelled to pay damages or serve sentences.[99] When these conditions are met, interventions targeting applications and users can be highly effective because they address harms where they occur.

But many AI-related harms involve actors who are difficult or impossible to reach through traditional enforcement: foreign adversaries, anonymous bad actors, autonomous AI systems, or diffuse harms for which no single defendant is responsible. In these situations, compute governance (Stages 1–2) becomes more valuable because it operates

---

[96] BILL DREXEL & CALEB WITHERS, CTR. FOR A NEW AM. SEC., AI AND THE EVOLUTION OF BIOLOGICAL NATIONAL SECURITY RISKS 26 (2024), available at https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AIBiologicalRisk_2024_Final.pdf.

[97] Sophie Williams, Jonas Schuett & Markus Anderljung, *On Regulating Downstream AI Developers*, 17 EUR. J. RISK REGUL. 94, 113–114 (2025), https://doi.org/10.1017/err.2025.10020.

[98] *Id.* at 107; IAN BROWN, ADA LOVELACE INST, ALLOCATING ACCOUNTABILITY IN AI SUPPLY CHAINS: A UK-CENTRED REGULATORY PERSPECTIVE 36–37 (2023), https://www.adalovelaceinstitute.org/resource/ai-supply-chains/.

[99] Sastry et al., *supra* note 86, at 32–39.

through chokepoints—concentrated infrastructure that can be controlled even when end users cannot be.[100]

Policymakers assessing enforcement feasibility should consider three questions. First, can compliance be observed?[101] Some requirements are relatively easy to verify (e.g., whether a model has been submitted for pre-deployment review, whether a chip shipment crossed a border). Others are harder to monitor (e.g., whether a company's internal safety practices match its public commitments). When compliance is difficult to observe, enforcement depends on whistleblowers, audits, or investigations, which are more resource-intensive yet less complete.[102]

Second, who will enforce? Regulatory requirements need an agency with the statutory authority, technical expertise, and budget to monitor compliance and pursue violators.[103] If that agency doesn't exist or is underfunded, requirements may go unenforced.

Third, what are the consequences for non-compliance, and are they sufficient to deter? A small fine may be treated as a cost of doing business; a large fine may deter only if the probability of detection is meaningful. An important note is that the certainty of apprehension and punishment matters far more than the severity of punishment for deterrence.[104] For AI governance, this implies that investing in monitoring and detection infrastructure may be more valuable than increasing statutory penalties.[105] It also suggests that highly publicized enforcement actions, which increase the perceived certainty of consequences, may matter more than their direct effects on the individual defendants involved.

---

[100] Lennart Heim, Markus Anderljung, Emma Bluemke & Robert Trager, *Computing Power and the Governance of AI*, GOVAI (Feb. 14, 2024), https://www.governance.ai/analysis/computing-power-and-the-governance-of-ai; Ananthi Al Ramiah et al., Toward a Global Regime for Compute Governance: Building the Pause Button 5 (June 25, 2025), https://arxiv.org/pdf/2506.20530; Erich Grunewald, *Compute Is a Strategic Resource*, INST. FOR AI POL'Y & STRATEGY (Sept. 2, 2025), https://www.iaps.ai/research/compute-is-a-strategic-resource.

[101] Tobin South, Private, Verifiable, and Auditable AI Systems 18 (Apr. 27, 2025) (PhD thesis, Massachusetts Institute of Technology), https://arxiv.org/abs/2509.00085 (discussing the tradeoff between user privacy and the need to ensure models are systematically verifiable and observable for regulatory compliance).

[102] *See* William Walter Finch & Marya Butt, *Gaps in AI-Compliant Complementary Governance Frameworks' Suitability (for Low-Capacity Actors), and Structural Asymmetries (in the Compliance Ecosystem)—A Systematic Review*, 5 J. CYBERSECURITY & PRIV. 101 (2025), https://www.mdpi.com/2624-800X/5/4/101 (observing that compliance becomes particularly difficult when institutional actors lack internal visibility into systems and lack resources for more involved verification).

[103] *See* Smith, *supra* note 60 (noting that a body "empowered" to administer a licensing regime will require technical expertise and sufficient financial and computational resources).

[104] Daniel S. Nagin, *Deterrence in the Twenty-First Century*, 42 CRIME & JUST. 199, 201–02 (2013).

[105] Justin Lynch & Emma Morrison, *Deterrence Through AI-Enabled Detection and Attribution*, HENRY A. KISSINGER CTR. FOR GLOB. AFFS. (July 2023), https://kissinger.sais.jhu.edu/programs-and-projects/kissinger-center-papers/deterrence-through-ai-enabled-detection-and-attribution/.

Market structure also shapes which tools are feasible. When an industry is concentrated, regulation is more tractable: fewer entities to monitor, greater compliance capacity, and reputational stakes that give enforcement actions bite.[106] When an industry is fragmented, direct regulation may be impractical, pushing policymakers toward upstream chokepoints or tools like liability that don't require entity-by-entity oversight. When entry barriers are low, regulations binding only incumbents may simply shift activity to less scrupulous providers.

6. Allocating responsibility to the least-cost avoider

A foundational principle of tort law holds that liability for a harm should be assigned to the party who can most cheaply and effectively prevent it: the "least-cost avoider."[107] This principle provides useful guidance for allocating responsibility across the AI ecosystem.[108]

In practice, identifying the least-cost avoider is contested. Consider an AI system that generates plausible-sounding medical misinformation that a user then relies upon.[109] Is the least-cost avoider the developer (who could have trained the model to be more cautious about medical claims), the deployer (who could have added guardrails or warnings), the platform that distributed the content (who could have labeled or filtered it), or the user who relied on it without verification? Each could have prevented the harm at some cost, and reasonable people will disagree about which intervention was cheapest or most effective (not to mention the contested normative question of who "ought" to have prevented the wrong).

The principle nevertheless provides a useful starting point: for harms resulting from unpredictable system failures (such as hallucinations), developers and deployers may be best positioned to invest in prevention because they are best positioned to build the scaffolding necessary to protect an AI system. For harms resulting from intentional misuse by end users, this principle may suggest focusing on the end user because they choose

---

[106] *See* Vipra & Korinek, *supra* note 83, at 33–37 (noting that frontier AI companies could be regulated under a similar model to public utilities if their role in the economy significantly expands because they operate in a market that tends towards natural monopoly or oligopoly).

[107] Paul Rosenzweig, *Content Moderation and the Least Cost Avoider* 2–3 (Am. U. Wash. C. of L. Joint PIJIP/TLS Rsch. Paper Series, Paper no. 125, 2024), https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1127&context=research.

[108] Maarten Herbosch, *How Existing Liability Frameworks Can Handle Agentic AI Harms*, LAWFARE (Dec. 3, 2025, 10:13 AM), https://www.lawfaremedia.org/article/how-existing-liability-frameworks-can-handle-agentic-ai-harms.

[109] *See* W. Nicholson Price II & I. Glenn Cohen, *Locating Liability for Medical AI*, 73 DEPAUL L. REV. 339, 341 (2024) (When a medical AI causes harm, "[n]either [the developer or the hospital] is precisely in position to be the cheapest cost avoider.").

whether and how to deploy AI for harmful purposes.[110] For harms that are harmful regardless of context (like CSAM generation), joint and several liability across the production chain may be appropriate to ensure all parties have incentives to prevent it.[111]

7. Is regulation the right tool?

Finally, policymakers should ask whether regulation is the right intervention at all. It is one tool among many, and depending on the type of harm, other interventions may be more effective.

Consider the range of alternatives. Investing in and adopting defensive technologies (like the kind described in subsection 2), for example, may be more effective in reducing cybersecurity vulnerabilities than regulation mandating cybersecurity best practices. Technical standards development, whether led by industry or the government, can also establish shared benchmarks for safety, interoperability, and performance that influence behavior.[112] Government procurement power can shape markets as companies compete on the metrics set by federal agencies for what they will purchase.[113] Antitrust enforcement (with remedies like divestiture) may address some concentration of power problems that regulation does not. And as consumers grow more sophisticated in evaluating AI products, market pressure alone may discipline some harms without the need for regulatory intervention.

Enforceability should weigh heavily in this assessment (discussed in subsection 5). Different tools require different enforcement capacities, and an intervention that cannot be meaningfully enforced may be worse than no intervention at all, creating the illusion of oversight while harmful practices continue or burdening compliant actors while bad actors ignore requirements.

---

[110] Jack Solowey, *AI Providers Should Not Be Liable for Users' Securities Violations*, CATO INST. (Apr. 4, 2024), https://www.cato.org/commentary/ai-providers-should-not-be-liable-users-securities-violations.

[111] FUTURE OF LIFE INST., TURNING VISION INTO ACTION: IMPLEMENTING THE SENATE AI ROADMAP 10 (2024), https://futureoflife.org/document/vision-into-action-senate-ai-roadmap/ (discussing how joint and several liability can reduce enforcement gaps).

[112] INT'L CHAMBER OF COM., ICC POLICY PAPER ON AI GOVERNANCE AND STANDARDS 5 (2025), https://iccwbo.org/wp-content/uploads/sites/3/2025/07/2025-ICC-Policy-Paper-AI-governance-and-standards.pdf.

[113] Nari Johnson et al., *Legacy Procurement Practices Shape How U.S. Cities Govern AI: Understanding Government Employees' Practices, Challenges, and Needs*, *in* PROCEEDINGS OF THE 2025 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 772, 774 (2025), https://dl.acm.org/doi/proceedings/10.1145/3715275; Giovanni Fabio Licata, *Transformative Public Procurement of Artificial Intelligence*, 14 LAWS 97 (2025), https://www.mdpi.com/2075-471X/14/6/97.

Ultimately, the question is both "which actor should we target?" and "what kind of intervention is most likely to work?" Policymakers should not rush past these questions by assuming regulation is the solution to every problem.

III. The Where: Which actor in the AI ecosystem does the intervention target?

The path from silicon chip to real-world harm is filled with actors that could prevent those harms. This primer organizes actors in the AI ecosystem into seven stages based on their role in the AI ecosystem: (1) chip designers and manufacturers, (2) cloud compute providers, (3) data suppliers, (4) model developers, (5) application deployers, (6) complementary and enabling platforms, and (7) end users. Each set of actors offers a potential leverage point for intervention.

These stages do not map neatly onto corporate structures: a single company can span multiple categories. Google, for example, functions as a cloud provider (Google Cloud), data supplier (via YouTube), model developer (Google DeepMind), and application deployer (Gemini). The purpose of this framework is to help policymakers regulate by the role an entity plays in the causal chain from the creation of an AI system to the materialization of harm, regardless of who performs each function.

Figure 1: Description of the actors in the AI ecosystem.

Stage 1: Chip Designers and Manufacturers

The semiconductor supply chain is a key intervention point for shaping AI's development and deployment. Training and running advanced AI systems requires specialized semiconductors like Application-Specific Integrated Circuits ("ASICs") or Graphics Processing Units ("GPUs"). These chips are designed by a handful of companies (mostly NVIDIA, AMD, Google) and produced by an even smaller number of foundries (mostly TSMC) using manufacturing equipment (extreme ultraviolet ("EUV") lithography machines) from a single supplier (ASML). This market concentration creates chokepoints that policymakers can leverage for policy enforcement.[114]

Examples of Policy Interventions

Export controls on AI chips. The U.S. government's primary tool for shaping international AI development thus far has been export controls on advanced semiconductors, first announced in late 2022 and updated several times since.[115] This policy is in flux as the Trump administration has flip-flopped on export controls for advanced AI chips to China. The most recent announcement was that NVIDIA can sell its advanced H200s to China.[116]

Compute infrastructure disclosure requirements. The Biden Administration's 2023 AI executive order (since rescinded) required entities acquiring or possessing large-scale computing clusters to report their existence, location, and total computing capacity to the government.[117] This was intended to give policymakers visibility into the physical infrastructure being assembled for advanced AI training.

---

[114] Sastry et al., *supra* note 86, at 28–30; Lennart Heim, *Compute and the Governance of AI - Talk*, HEIM.XYZ (Nov. 5, 2023), https://blog.heim.xyz/compute-and-the-governance-of-ai-talk/.

[115] *Fact Sheet: Ensuring U.S. Security and Economic Strength in the Age of Artificial Intelligence*, THE WHITE HOUSE (Jan. 13, 2025), https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2025/01/13/fact-sheet-ensuring-u-s-security-and-economic-strength-in-the-age-of-artificial-intelligence/ (summarizing Biden-era export control policy); Press Release, Bureau of Indus. & Sec., Department of Commerce Announces Rescission of Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls (May 13, 2025), https://www.bis.gov/press-release/department-commerce-announces-rescission-biden-era-artificial-intelligence-diffusion-rule-strengthens.

[116] Rogé Karma, *Trump is Throwing Away America's AI Dominance*, ATLANTIC (Dec. 12, 2025), https://www.theatlantic.com/economy/2025/12/trumps-china-ai-chips/685235/; Chris McGuire, China's AI Chip Deficit: Why Huawei Can't Catch Nvidia and U.S. Export Controls Should Remain, COUNCIL ON FOREIGN REL. (Dec. 15, 2025, 9:19 AM), https://www.cfr.org/article/chinas-ai-chip-deficit-why-huawei-cant-catch-nvidia-and-us-export-controls-should-remain; Alasdair Phillips-Robins, *Don't Panic Yet Over AI Chip Sales to China*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Dec. 12, 2025), https://carnegieendowment.org/emissary/2025/12/china-ai-chip-sales-nvidia-trump?lang=en.

[117] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Nov. 1, 2023), *repealed by* Exec. Order No. 14,148, § 2(ggg), 90 Fed. Reg. 8,237, 8,240 (Jan. 20, 2025).

On-chip governance mechanisms. Researchers have proposed adding technical components to chips that would give regulators visibility into large training runs without compromising AI companies' commercial interests.[118] Though requiring further research into technical feasibility, verification mechanisms have been a prerequisite for international cooperation in other domains (like nuclear nonproliferation) and may be just as valuable for AI governance.[119]

Advantages of Targeting This Stage

Focusing policy interventions on semiconductor companies has four key advantages: excludability, quantifiability, detectability, and enforcement feasibility.[120]

First, computing resources are excludable because people can be prevented from accessing them. Unlike data or algorithms, which are easily copied and shared, there are a finite number of operations that a single GPU can perform. If one actor is fully utilizing those operations, no one else can.

Second, compute is quantifiable—it "can be easily measured, reported, and verified" in terms of the operations per second a chip can perform or its communication bandwidth with other chips.[121]

Third, large-scale training runs are detectable. The most advanced AI models currently require large-scale training runs using thousands of specialized chips concentrated in power-intensive data centers. These facilities are detectable by third parties, with some visible from satellite imagery.

Finally, the semiconductor supply chain is extraordinarily concentrated, making enforcement feasible.[122] NVIDIA controls 80-95% of the market for AI chip design; TSMC fabricates approximately 90% of advanced chips; ASML supplies 100% of the extreme ultraviolet lithography machines used by leading foundries. This concentration makes the other three advantages actionable: fewer actors makes monitoring and enforcement easier.[123]

---

[118] Onne Aarne, Tim Fist & Caleb Withers, *Secure, Governable Chips*, CTR. FOR A NEW AM. SEC. (Jan. 8, 2024), https://www.cnas.org/publications/reports/secure-governable-chips.

[119] Robert F. Trager et al., International Governance of Civilian AI 18–19 (Oxford Martin AI Governance Initiative Whitepaper, Aug. 2023), https://cdn.governance.ai/International_Governance_of_Civilian_AI_OMS.pdf.

[120] For a more detailed analysis, *see* Sastry et al., *supra* note 86; Heim, *supra* note 113. This section draws heavily on their excellent work.

[121] Sastry et al., *supra* note 86, at 4.

[122] SAIF M. KHAN, DAHLIA PETERSON & ALEXANDER MANN, THE SEMICONDUCTOR SUPPLY CHAIN: ASSESSING NATIONAL COMPETITIVENESS, CTR. FOR SEC. & EMERGING TECH. 22–23, 30–31 (2021), https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/.

[123] CHRIS MILLER, CHIP WAR: THE FIGHT FOR THE WORLD'S MOST CRITICAL TECHNOLOGY 23–25 (2022).

Disadvantages of Targeting This Stage

Despite these advantages, semiconductor-focused interventions have three significant drawbacks: they rely on an assumption that may not hold, they are blunt and they become less effective the more they are used.

First, semiconductor-focused regulation relies on compute serving as an effective proxy for capability. If AI performance improves such that less-than-state-of-the-art models become capable of serious harm, or if algorithmic efficiency gains reduce the compute needed for dangerous capabilities, then semiconductor-focused interventions will miss their intended target.[124] Additionally, as Arvind Narayanan and Sayash Kapoor have argued, "AI safety is not a model property," since whether an output is harmful depends on context, and capability restrictions are thus, although laudable, a misguided approach.[125]

Chip-level interventions are also among the bluntest available.[126] Because semiconductors are general-purpose inputs, restrictions at this stage cannot distinguish between beneficial and harmful uses of AI. Export controls that limit access to advanced chips, for example, constrain medical research and climate modeling just as much as they constrain weapons development or surveillance. Policymakers targeting this stage risk throttling AI development altogether rather than surgically addressing specific harms.

Finally, semiconductor interventions are valuable in the short term but potentially counterproductive in the long-term. On-chip governance mechanisms and export controls, while potentially controlling how chips are used or limiting adversaries' access to them in the short term, encourage investment in alternatives. On-chip restrictions could push buyers toward more open alternatives produced in other jurisdictions. Some have argued U.S. export controls benefit China's semiconductor industry by encouraging further investment,[127] though others believe such indigenization fears are overblown.[128] And

---

[124] Venkat Somala, Anson Ho & Séb Krier, *Three Challenges Facing Compute-Based AI Policies*, EPOCH AI (Sept. 11, 2025), https://epoch.ai/gradient-updates/three-issues-undermining-compute-based-ai-policies.

[125] Arvind Narayanan & Sayash Kapoor, *AI Safety is Not a Model Property*, AI AS NORMAL TECH. (Mar. 12, 2024), https://www.normaltech.ai/p/ai-safety-is-not-a-model-property; see also Andrew Ng, *The Problem With California's AI Bill*, TIME (Aug. 29, 2024, 9:42 AM), https://time.com/7016134/california-sb-1047-ai/ (arguing that the applications of general purpose technologies such as AI should be regulated, rather than the technologies themselves).

[126] *See* Part II.4, *supra*.

[127] *See, e.g.*, Sujai Shivakumar, Charles Wessner & Thomas Howell, *The Limits of Chip Export Controls in Meeting the China Challenge*, CTR. FOR STRATEGIC & INT'L STUD. (Apr. 14, 2025), https://www.csis.org/analysis/limits-chip-export-controls-meeting-china-challenge; Rodrigo Balbontin, *Backfire: Export Controls Helped Huawei and Hurt U.S. Firms*, INFO. TECH. & INNOVATION FOUND. (Oct. 27, 2025), https://itif.org/publications/2025/10/27/backfire-export-controls-helped-huawei-and-hurt-us-firms/.

[128] *Stop Selling the Rope*, AM. COMPASS (Oct. 27, 2025), https://americancompass.org/stop-selling-the-rope/.

geopolitical competition may prevent countries from regulating their own semiconductor industries at all, for fear of ceding a competitive edge. Finally, compute-based restrictions may encourage innovation that renders compute less likely to be a bottleneck in the future. For example, DeepSeek's efficiency may be an unintended consequence of limiting Chinese firms' access to computing resources, with compute constraints forcing researchers to develop more efficient algorithms.

Summary

Intervention	Status	What It Does	Key Limitation
Export controls	In flux	Restricts chip sales to adversaries	Incentivizes alternatives
Compute infrastructure disclosure requirements	Rescinded	Requires reporting on existing, location, and capacity of large computing clusters	Relies on compute as proxy for capability
On-chip monitoring	Proposed	Hardware-level training verification	May push buyers elsewhere; Technical feasibility.

Semiconductor-based interventions are a powerful policy tool. They work best for harms that are identifiable without much context, where a single malicious actor's access to advanced capabilities increases risk regardless of how they're used. They are particularly valuable when traditional enforcement is impractical—against foreign actors, in situations involving diffuse harms, or when agencies cannot possibly monitor all end users. But policymakers should be aware that these interventions are a blunt instrument that incentivize workarounds, and they depend on compute remaining a bottleneck for dangerous capabilities.

Stage 2: Cloud Compute Providers

A handful of cloud computing providers—Amazon Web Services, Microsoft Azure, and Google Cloud—operate the computing infrastructure used by most developers to train and deploy AI models. Because this stage and the previous one both focus on the computational resources underpinning advanced AI, they share many characteristics and are often grouped together under the general heading of "compute governance." This section focuses on the ways in which targeting cloud providers differs from targeting the semiconductor supply chain.[129]

[129] *See generally* Lennart Heim et al., Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation (Oxford Martin Sch. Whitepaper, Mar. 13, 2024), https://www.oxfordmartin.ox.ac.uk/publications/governing-through-the-cloud-the-intermediary-role-of-compute-providers-in-ai-regulation (arguing that cloud compute providers should have legal and ethical

Examples of Policy Interventions

Know-Your-Customer (KYC) requirements. Efforts to impose KYC obligations on cloud providers date back to Trump's first-term EO 13984 (2021),[130] which Biden's EO 14110 (2023)[131] later expanded with AI-specific reporting mandates. Providers would be required to notify the government when foreign actors access computing resources above certain thresholds, aiming to prevent adversaries from anonymously using U.S. infrastructure for dangerous AI development.

Cybersecurity compliance and AI safety standards. In 2024, the Commerce Department's Bureau of Industry and Security (BIS) proposed rules requiring cloud providers and their clients to report on frontier AI development activities, including compliance with cybersecurity frameworks and results of mandatory red-teaming tests.[132] This would leverage cloud providers as enforcers to ensure hosted AI projects meet security benchmarks.[133]

Government oversight of frontier AI training. Some researchers have recommended requiring licenses for AI training that exceeds certain compute thresholds.[134] Cloud providers would require government approval before allowing training runs capable of producing frontier models—resembling licensing in the aerospace and nuclear energy industries.

Advantages of Targeting This Stage

Cloud computing shares the same fundamental advantages as semiconductor governance—concentration, excludability, quantifiability, and detectability—but offers additional benefits for regulators.

First, cloud providers' business models already require extensive monitoring. Because they charge customers based on usage, they precisely measure resource consumption through metrics like floating-point operations, GPU hours, and energy

responsibilities associated with AI deployment); *see also* Janet Egan & Lennart Heim, *America Should Rent, Not Sell, AI Chips to China*, RAND (Aug. 15, 2025), https://www.rand.org/pubs/commentary/2025/08/america-should-rent-not-sell-ai-chips-to-china.html (arguing that cloud services offer greater control over how Chinese firms use chips compared to sales).

[130] Exec. Order No. 13,984, 86 Fed. Reg. 6,837 (Jan. 19, 2021).

[131] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Nov. 1, 2023).

[132] Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, 89 Fed. Reg. 73,612 (proposed Sept. 11, 2024) (to be codified at 15 C.F.R. pt. 702).

[133] Markus Anderljung et al., Frontier AI Regulation: Managing Emerging Risks to Public Safety (Nov. 7, 2023), https://arxiv.org/abs/2307.03718; Haydn Belfield, Domestic Frontier AI Regulation, an IAEA for AI, an NPT for AI, and a US-led Allied Public-Private Partnership for AI: Four Institutions for Governing and Developing Frontier AI 7 (July 8, 2025), https://arxiv.org/abs/2507.06379.

[134] *See* Smith, *supra* note 60.

consumption. This existing infrastructure can be repurposed for regulatory compliance with less added burden.

In addition, unlike the semiconductor supply chain, which is distributed across the United States' allied nations, including the Netherlands, Japan, and Germany, the world's largest cloud providers are headquartered in the United States. This gives American policymakers more jurisdictional flexibility in designing and enforcing regulations.[135]

Cloud providers can also respond to non-compliance immediately. Unlike semiconductor restrictions, where there is a lag between placement on an export control list and actual business disruption, cloud access can be revoked in real time (similar to how banks can freeze accounts). This makes cloud-based enforcement more like financial regulation than trade regulation.

Disadvantages of Targeting This Stage

Cloud provider interventions share the two core weaknesses of semiconductor controls and add a third.

Like semiconductor interventions, cloud-based regulation relies on the assumption that advanced AI requires access to large compute clusters. If algorithmic efficiency improves enough that dangerous capabilities can be achieved with modest resources, then cloud-focused controls will miss their target.

Cloud restrictions also risk sparking the same kind of workarounds and indigenization that undermine export controls. Developers facing extensive compliance requirements may switch to less-regulated providers in other jurisdictions, and countries seeking AI independence or "sovereignty"[136] may invest in domestic cloud infrastructure precisely to escape U.S.-based oversight. The more effective cloud controls are in the short term, the stronger the incentive to route around them in the long term.

Finally, cloud interventions also raise concerns that detailed tracking of activity on clients' servers could expose information that companies treat as trade secrets like training techniques.[137] And broad authority to cut off companies' access to computing resources

---

[135] *See* Janet Egan & Lennart Heim, Oversight for Frontier AI Through a Know-Your-Customer Scheme for Compute Providers 3–4 (Oct. 20, 2023), https://arxiv.org/abs/2310.13625 (arguing that regulating digital access to compute "offers more precise controls" than chip export controls and that the U.S., as a dominant compute provider, "can exert broad influence through a domestically implemented scheme")

[136] Swati Srivastava & Justin Bullock, AI, Global Governance, and Digital Sovereignty 10–11 (Oct. 24, 2024), https://arxiv.org/abs/2410.17481; David Wood, et al., *Sovereign AI: Own Your AI Future*, ACCENTURE (last accessed Mar. 2, 2026), https://www.accenture.com/content/dam/accenture/final/accenture-com/document-4/Sovereign-AI-Report.pdf (advising private firms on how to address AI sovereignty issues).

[137] Sastry et al., *supra* note 86, at 61.

creates potential for abuse, as officials might target companies disfavored by the governing party or use access as leverage for purposes beyond AI safety.

Summary

Intervention	Status	What It Does	Key Limitation
KYC requirements	In flux	Verify identity of foreign renters	Developers may shift to non-U.S. providers; Privacy concerns
Cybersecurity standards	Proposed	Require red-teaming, security compliance	Potential for government abuse; Monitoring may expose trade secrets
Training run licensing	Speculative	Approval for frontier training	Relies on compute as proxy for capability

Cloud provider interventions function as a more responsive version of semiconductor controls—enforcement can happen immediately rather than through supply-chain disruption. They are particularly valuable when real-time response matters or when tracking the identity of AI developers (rather than just their capability) is important. However, they come with greater privacy trade-offs and the same diminishing-returns problem as other compute governance approaches.

Stage 3: Data Suppliers

AI models are trained on data sourced from many entities: large-scale web scrapers (like C4), crowdsourced platforms (like Amazon Mechanical Turk), and specialized data labeling companies (like ScaleAI, Mercor).[138] The goal of data-supplier-focused interventions is to improve the upstream collection and quality of AI training data to reduce downstream harms.[139]

Examples of Policy Interventions

Individual data rights. Several states require data brokers to register with the government and disclose details about the data they collect, how it's used, and their sharing practices. California's Delete Act aims to allow residents to request deletion of their

[138] *See* Karyna Naminas, *AI Training Data: Top Sources and Dataset Providers*, LABEL YOUR DATA (Nov. 13, 2025), https://labelyourdata.com/articles/machine-learning/ai-training-data (describing categories and sources of AI training data).

[139] Araz Taeihagh, *Generative AI Governance Challenges*, 44 POL'Y & SOC'Y 1, 3–5 (2025), https://academic.oup.com/policyandsociety/article/44/1/1/7997395.

personal information from data brokers through a single request.[140] Illinois's Biometric Information Privacy Act (BIPA) requires opt-in consent for entities collecting biometric data,[141] significantly affecting AI facial recognition training. With AI, honoring these rights may involve removing data from supplier databases, retraining models, or filtering outputs traceable to deleted data, though the FTC has ordered complete model deletion in cases of egregious data misuse.[142]

Collective licensing and compensation mechanisms. John Axhamn has proposed an "Extended Collective Licensing (ECL)" scheme that would allow AI developers to legally train on copyrighted works while ensuring rightsholders receive compensation.[143] A Congressional Research Service report discusses these frameworks as potential solutions to the tension between AI training and intellectual property rights.[144]

Public datasets. Governments and research institutions could create and maintain high-quality public datasets. The Trump administration's Genesis Mission is an attempt to unify federal scientific archives into a centralized, standardized platform, democratizing

---

[140] Act of Oct. 10, 2023, ch. 709, 2023 Cal. Stat. (S.B. 362); *see also Accessible Deletion Mechanism – Delete Request and Opt–out Platform ("DROP") System Requirements*, CAL. PRIV. & PROT. AGENCY (last accessed Mar. 2, 2026), https://cppa.ca.gov/regulations/drop.html (summarizing the legislation).

[141] Brett M. Doran, Jena M. Valdetero & Zachary Pestine, *BIPA Update: Illinois Limits Liability and Clarifies Electronic Consent for Biometric Data Collection*, GREENBERG TRAURIG (Aug. 14, 2024), https://www.gtlaw.com/en/insights/2024/8/bipa-update-illinois-limits-liability-and-clarifies-electronic-consent-for-biometric-data-collection#:~:text=The%20Illinois%20General%20Assembly%20enacted,of%20hand%20or%20facial%20geometry.

[142] Zachary Sorenson, *Everalbum, Inc: In First Facial Recognition Misuse Settlement, FTC Requires Destruction of Algorithms Trained on Deceptively Obtained Photos*, JOLT Digest (Jan 13, 2021), https://jolt.law.harvard.edu/digest/everalbum-inc-in-first-facial-recognition-misuse-settlement-ftc-requires-destruction-of-algorithms-trained-on-deceptively-obtained-photo (analyzing FTC settlement requiring model deletion when training data included deceptively obtained photographs); Lauren Merk & Bailey Sanchez, *FTC Requires Algorithmic Disgorgement as COPPA Remedy for First Time*, FUTURE OF PRIV. F. (Mar. 14, 2022), https://fpf.org/blog/ftc-requires-algorithmic-disgorgement-as-a-coppa-remedy-for-first-time (describing the FTC's use of the "algorithmic disgorgement" remedy).

[143] For more detail, *see* Part II of Johan Axhamn, *Extended Collective Licensing for Use of Copyrighted Works for Machine Learning*, 48 COLUM. J. L. & ARTS 523 (2025), https://doi.org/10.52214/jla.v48i4.13933; *see also* David W. Opderbeck, *Copyright in AI Training Data: A Human-Centered Approach*, 76 OKLA. L. REV. 952, 996–99 (2024) (discussing a collective rights approach to licensing training data).

[144] CHRISTOPHER T. ZIRPOLI, CONG. RSCH. SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW 6 (2025), https://www.congress.gov/crs-product/LSB10922 ("One scholar, acknowledging that '[i]t would . . . be impossible for an AI developer to identify and clear billions of rights claims on an individual basis,' argues that it may be feasible instead to create markets for AI training data via means such as . . . collective management organizations (or CMOs, such as those that manage rights to musical works)").

AI development for smaller companies and academic researchers who lack resources to develop proprietary datasets.[145]

Advantages of Targeting This Stage

Addressing data quality and legality at the point of collection and aggregation can prevent problems from multiplying as models are deployed.[146]

Many data supplier regulations align with existing privacy laws like GDPR and CCPA.[147] This allows policymakers to build on established definitions, enforcement mechanisms, and compliance infrastructure rather than creating entirely new regulatory regimes.

The data broker[148] and AI data labeling[149] industries are also relatively concentrated, allowing authorities to oversee a significant portion of data flowing into AI systems by regulating a manageable number of entities.

Disadvantages of Targeting This Stage

Data suppliers can operate globally, potentially sourcing data from or locating servers in jurisdictions with weaker regulations. AI developers might bypass rules by using offshore brokers or scraping data outside the state's reach.[150]

---

[145] *Fact Sheet: President Donald J. Trump Unveils the Genesis Mission to Accelerate AI for Scientific Discovery*, THE WHITE HOUSE (Nov. 24, 2025), https://www.whitehouse.gov/fact-sheets/2025/11/fact-sheet-president-donald-j-trump-unveils-the-genesis-missionto-accelerate-ai-for-scientific-discovery/.

[146] *See* Nithya Sambasivan et al., *"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI*, *in* PROCEEDINGS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (Article No. 39, 2021), https://dl.acm.org/doi/10.1145/3411764.3445518 (defining, identifying, and presenting evidence on "data cascades" based on practitioner interviews).

[147] EUR. PARLIAMENTARY RSCH. SERV., THE IMPACT OF THE GENERAL DATA PROTECTION REGULATION (GDPR) ON ARTIFICIAL INTELLIGENCE 76 (2020), https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf (finding no incompatibility between the GDPR, AI, and big data).

[148] *See* Laura Abrardi, Carlo Cambini & Flavio Pino, *Data Brokers Competition, Synergic Datasets, and Endogenous Information Value*, 103 INT'L J. INDUS. ORG. 103146, 103149 (2025) ("The data brokerage ecosystem generates revenues of over USD 200 billion a year and it is dominated by a few very large companies").

[149] *Surge AI*, SACRA (last accessed Mar. 8, 2026), https://sacra.com/c/surge-ai/ (describing a duopoly in which Scale AI and Surge AI together serve the roughly twelve frontier labs — including OpenAI, Google, Anthropic, Microsoft, and Meta — that drive demand for expert AI training data).

[150] Sotiris Spyrou, *Regulatory Arbitrage in the AI Era: Why Compliance Complexity Demands Independent Validation*, VERIFYAI (July 30, 2025), https://verityai.co/blog/regulatory-arbitrage-ai-compliance.

Defining who qualifies as a regulated "data supplier" is also difficult.[151] Rules could be too narrow (missing web scrapers) or too broad (burdening nonprofits like Wikipedia or individual bloggers). And compliance costs may further consolidate the market around larger players who can afford them.

Finally, regulations restricting collection, sale, or use of data may face First Amendment challenges. The Supreme Court's decision in Sorrell v. IMS Health held that the "creation and dissemination of information are speech," meaning data regulations can potentially trigger heightened constitutional scrutiny.[152]

Summary

Intervention	Status	What It Does	Key Limitation
Consent requirements	In effect (BIPA)	Opt-in for sensitive data	Narrow scope (biometric data only)
Collective licensing	Proposed	Training with creator compensation	Complex administration
Public data commons	Proposed	Government-curated datasets	Funding and maintenance

Data supplier regulations are most useful for harms directly linked to data inputs—privacy violations, copyright infringement, and biased training data—and where key data sources are concentrated and identifiable. They also benefit from alignment with existing privacy frameworks like GDPR and CCPA. They are less effective when data is highly diffuse, when suppliers operate across jurisdictions, or when defining the scope of regulated entities is difficult.

Stage 4: AI Model Developers

This stage focuses on organizations that research, develop, and train large-scale AI models—entities like OpenAI, Anthropic, Google DeepMind, Meta AI, and xAI. Given their central role in developing core AI capabilities, model developers are frequent targets for policy intervention.[153]

---

[151] Justin Sherman, *Federal Privacy Rules Must get "Data Broker" Definitions Right*, LAWFARE (Apr. 8, 2021, 11:00 AM), https://www.lawfaremedia.org/article/federal-privacy-rules-must-get-data-broker-definitions-right.

[152] 564 U.S. 552, 570 (2011).

[153] *See, e.g.*, Dean W. Ball & Ketan Ramakrishnan, *Entity-Based Regulation in Frontier AI Governance*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 7, 2025), https://carnegieendowment.org/russia-eurasia/research/2025/07/artificial-intelligence-regulation-united-states (supporting regulation of frontier model developers).

Examples of Policy Interventions

Transparency and Disclosure Mandates. The most common type of intervention at the state level promotes transparency through disclosure requirements.[154] These can take several forms: requiring companies to publish reports on model development and testing processes; mandating watermarks or content provenance techniques to label AI-generated content; creating whistleblower protections for AI lab employees; and establishing incident reporting requirements for significant safety events.

Testing, Design, and Oversight Requirements. These interventions focus on ensuring safety through design mandates and ongoing oversight.[155] They might include ex ante design features (cybersecurity safeguards, content filtering, "kill switches"), standardized pre-deployment evaluations, red-teaming by internal safety experts,[156] or third-party audits by civil society organizations (like METR) or government agencies (like the UK AISI or US CAISI).[157]

Liability frameworks. Policy researchers have advocated holding model developers liable for harms caused by their systems.[158] California's SB 53, for example, authorizes the state Attorney General to bring civil actions against developers for failure to report safety incidents or comply with their own frameworks, with damages up to $1,000,000.[159] Insurance requirements could complement direct liability by ensuring compensation is available and creating actuarial signals about the riskiest models.[160] For catastrophic risks, policymakers could consider ex ante mechanisms like mandatory

---

[154] *See, e.g.*, Lam Tran, *Governing Frontier AI: California's SB 53*, LAWFARE (Oct. 21, 2025, 10:16 AM), https://www.lawfaremedia.org/article/governing-frontier-ai--california-s-sb-53 (discussing transparency and disclosure requirements in California).

[155] Anderljung et al., *supra* note 132, at 23–28.

[156] *See, e.g.*, Deep Ganguli et al., Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned (Sept. 16, 2022), https://arxiv.org/abs/2209.07858 (describing red-teaming methods and challenges).

[157] Shayne Longpre et al., A Safe Harbor for AI Evaluation and Red Teaming (Mar. 7, 2024), https://arxiv.org/abs/2403.04893 (third-party audits by civil society organizations); *Early Lessons From Evaluating Frontier AI Systems*, AI SEC. INST. (UK AISI) (Oct. 24, 2024), https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems (audits by the UK AISI).

[158] Weil, *supra* note 66; Dean W. Ball, *How Should AI Liability Work? (Part I)*, HYPERDIMENSIONAL (Feb. 20, 2025), https://www.hyperdimensional.co/p/how-should-ai-liability-work-part; Dean W. Ball, *How Should AI Liability Work? (Part II)*, HYPERDIMENSIONAL (Feb. 26, 2025), https://www.hyperdimensional.co/p/how-should-ai-liability-work-part-3df.

[159] 2025 Cal. Stat. ch. 138 (S.B. 53), available at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.

[160] Gabriel Weil et al., *Insuring Emerging Risks from AI* (Inst. for L. & A.I., Working Paper No. 4-2024, 2024), https://law-ai.org/insuring-emerging-risks-from-ai/.

liability insurance or industry-wide compensation funds, modeled on frameworks like the Price-Anderson Act for nuclear accidents.[161]

Advantages of Targeting this Stage

Transparency measures are among the least controversial categories of model developer regulation.[162] Even experts who disagree about the speed and shape of AI risks often agree on the need for more transparency.[163] The rationale is that disclosure empowers users, researchers, and regulators to make informed decisions—if people know how a model was created and tested, they can better assess its reliability and risk.

Testing and oversight requirements can target vulnerabilities in base models that propagate to downstream applications. If a foundation model has a security vulnerability or can be tricked into revealing training data, applications built on top of it can inherit those vulnerabilities.[164] It also leverages technical expertise outside government: regulators can set standards and review results while outsourcing the highly technical evaluation process to specialists.[165]

Finally, liability incentivizes developers to prevent harms by requiring them to pay for damage caused by their models. Liability forces companies to internalize externalities rather than shifting costs to users or society. It also can function as a floor of legal protection, with some egregious AI harms likely falling within the scope of existing tort law.[166]

---

[161] Matthew van der Merwe, Ketan Ramakrishnan & Markus Anderljung, *Tort Law and Frontier AI Governance*, LAWFARE (May 24, 2024, 1:38 PM), https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance.

[162] Even some frontier AI labs have acknowledged the need for transparency measures. *See The Need for Transparency in Frontier AI*, ANTHROPIC (July 7, 2025), https://www.anthropic.com/news/the-need-for-transparency-in-frontier-ai.

[163] See Yoshua Bengio et al., Managing AI Risks in an Era of Rapid Progress 5 (Nov. 12, 2023), https://arxiv.org/abs/2310.17688 (warning that advanced AI could pose existential risks and calling on regulators to require "whistleblower protections, incident reporting, registration of key information on frontier AI systems and their data sets throughout their life cycle, and monitoring of model development and supercomputer usage"); Written Statement of Andrew Ng Before the U.S. Senate AI Insight Forum 2–6 (Dec. 6, 2023), https://www.schumer.senate.gov/imo/media/doc/Andrew%20Ng%20-%20Statement.pdf (dismissing AI extinction risk as "sensationalist" and vanishingly improbable, but urging Congress to "[m]andate AI [t]ransparency" by giving government agencies and academics "the ability to obtain relevant information from large platforms").

[164] Rishi Bommasani et al., On the Opportunities and Risks of Foundation Models 105–06 (Aug. 16, 2021), https://arxiv.org/abs/2108.07258.

[165] Gillian K. Hadfield & Jack Clark, *Regulatory Markets: The Future of AI Governance*, 65 JURIMETRICS J. 195, 197 (2026).

[166] *See, e.g.*, Kate Payne, *An AI Chatbot Pushed a Teen to Kill Himself, a Lawsuit Against its Creator Alleges*, ASSOCIATED PRESS, Oct. 25, 2024, https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0 (describing a lawsuit under existing tort doctrine against the makers of a chatbot following the suicide of a 14-year-old).

Disadvantages of Targeting this Stage

Transparency requirements carry their own risks. Detailed disclosures about training data or architectures could expose trade secrets, allowing competitors (including foreign actors) to free-ride on research investments. Watermarking faces technical feasibility concerns: determined actors can remove watermarks, and open-source models that don't apply them will remain unlabeled. And in the United States, compelled disclosure requirements may face First Amendment challenges as compelled speech.[167]

Oversight requirements face a different problem: defining "safe enough" is extremely difficult. AI systems can fail in unpredictable ways, and any set of evaluations risks missing novel failure modes. Worse, developers might learn to game required tests, tuning models to pass evaluations without truly improving safety.[168] Extensive testing requirements can also slow innovation and concentrate the market as larger, well-resourced companies absorb compliance costs more easily.

Finally, liability is limited by the fact that many AI harms are not within developers' sole control. AI systems are general-purpose tools used in countless contexts far removed from what creators envisioned. If an open-source model is integrated into hundreds of applications, it may be unfair to blame the model's creators for a particular application's failure or a user's misuse. Defining "reasonable care" for AI, determining causation when multiple actors are involved, and apportioning liability (joint and several vs. proportional) all present difficult legal questions.[169] Liability also functions best when wrongdoers can afford to pay; when the defendant lacks resources or the harm is too large, liability does not fully compensate victims and is a less effective deterrent.

---

[167] *See* Zauderer v. Off. of Disciplinary Couns. of Sup. Ct. of Ohio, 471 U.S. 626, 650–51 (1985) (noting that "[w]e have … held that in some instances compulsion to speak may be as violative of the First Amendment as prohibitions on speech" but declining to find that certain compelled disclosures in attorney advertising violated the First Amendment); *see also* Bahrad Sokhansanj, *xAI's Challenge to California's AI Training Data Transparency Law (AB2013)*, INST. FOR L. & AI (Jan. 2026), https://law-ai.org/xais-challenge-to-californias-ai-training-data-transparency-law-ab2013/.

[168] Nurit Cohen-Inger et al., Forget What you Know About LLMs Evaluations – LLMs are Like a Chameleon (Sept. 17, 2025), https://arxiv.org/abs/2502.07445v2.

[169] ISKANDAR HAYKEL, AM. FOR RESPONSIBLE INNOVATION, AI LIABILITY REPORT: THE STICK, THE CARROT, AND THE NET 8 (2025), https://ari.us/wp-content/uploads/2025/08/AI-Liability-Report-The-Stick-the-Carrot-and-the-Net.pdf.

Summary

Intervention	Status	What It Does	Key Limitation
Transparency reports	Emerging (state laws)	Require disclosure of training and testing	Trade secret exposure
Watermarking / provenance	Proposed	Label AI-generated content at model level	Technical feasibility; evasion
Whistleblower protections	Proposed	Shield employees reporting safety concerns	Enforcement difficulty
Pre-deployment testing	Emerging	Standardized safety evaluations before release	Defining "safe enough"; gaming
Third-party audits	Emerging	Independent evaluation by outside specialists	Cost; market concentration
Developer liability	Emerging (e.g., CA SB 53)	Civil actions for safety violations	Causation; general-purpose tools
Mandatory insurance	Proposed	Ensure compensation for catastrophic harms	Actuarial uncertainty

Model developers are a natural focal point for AI regulation because they control the foundational capabilities on which downstream applications depend. Transparency requirements enjoy the broadest support and face the lowest political barriers, though they must be designed to avoid exposing proprietary information. Testing and oversight requirements can catch vulnerabilities before they propagate downstream, but they must be iteratively adjusted as capabilities evolve—static evaluations will quickly become obsolete. Liability frameworks complete the picture by internalizing costs, but their effectiveness depends on resolving difficult questions about causation, apportionment, and the financial capacity of defendants to pay for harms their models enable.

Stage 5: Application Deployers

The application layer is where AI capabilities are operationalized into products and services. The entity developing the model can also be the deployer (OpenAI's ChatGPT is built on its GPT models), but different companies can also build applications on top of models developed by others. Applications span nearly every sector: predictive algorithms for employment screening (HireVue, Pymetrics) and criminal justice risk assessments (COMPAS); generative models powering chatbots (Character.ai) and coding environments (Cursor); and agentic systems capable of completing tasks without human intervention (Claude Code, Devin).

Examples of Policy Interventions

Incident reporting. When AI applications malfunction or cause harm, rapid disclosure enables regulators and the public to identify patterns and respond before problems spread. Analogous requirements exist across regulated industries: the FDA mandates adverse-event reporting for medical devices, NHTSA requires crash reporting for autonomous vehicles, and the FAA tracks near-miss incidents in aviation. Extending this model to AI deployers would require companies to notify a designated agency when their systems produce significant failures—whether a hiring algorithm systematically excludes qualified candidates or a medical diagnostic tool generates dangerous misdiagnoses. The core difficulty is defining what counts as a reportable "incident" for general-purpose AI systems that operate across diverse contexts.

Tort liability. Application deployers are also potential targets for liability.[170] Injured parties can sue the deployer for compensation if a particular application causes harm. One mother, for instance, sued Character.AI over her son's suicide, alleging design defects.[171] In a separate case, the parents of a sixteen-year-old sued OpenAI, alleging that ChatGPT acted as a "suicide coach" that encouraged their son's suicidal ideation.[172] The possibility of liability creates incentives to build safe products. Legislatures can play an important role in ensuring this system functions well. California's AB 316, for example, is a recently enacted bill that aims to ensure there is no AI exception to existing tort liability frameworks.[173] It prevents defendants from disclaiming responsibility by arguing that an AI agent acted autonomously.[174] Legislatures could also codify duties of care rather than leaving courts to define them through the common-law process.

Age restrictions. These provisions require deployers to implement age-appropriate design for systems used by minors, including limiting data collection, obtaining parental consent, and requiring plain-language explanations. Texas's SCOPE Act (2024) offers a comprehensive template: it requires parental consent before minors can create accounts, prohibits in-app purchases and geolocation tracking for minors, bans targeted advertising to children, and mandates content filtering for material promoting self-harm, substance abuse, or exploitation.[175] This approach is valuable because it shifts the compliance burden

---

[170] *See* John G. Browning, *Whose Bot Is It Anyway? Determining Liability for AI-Generated Content*, 45 N. ILL. UNIV. L. REV. 340 (2025) (analyzing emerging liability issues for developers and deployers of AI chatbots).

[171] Payne, *supra* note 167.

[172] Chatterjee, *supra* note 57.

[173] Act of Oct. 13, 2025, ch. 672, 2025 Cal. Stat. (codified at Cal. Civ. Code § 1714.46); *see also* Lior, *supra* note 66 (arguing existing tort doctrines should be applied to AI).

[174] Cal. Civ. Code § 1714.46.

[175] Tex. Bus. & Com. Code Ann. §§ 509.001–.152.

onto platforms rather than parents, creating structural protections that don't depend on individual digital literacy or vigilance.

Self-exclusion. These statutes require deployers to build protective features into products—such as session caps, break reminders, and disabling infinite scroll—that help users limit their own consumption. The UK Gambling Commission's "reality check" regulations provide a direct model:[176] operators must display recurring on-screen reminders of elapsed session time that users must actively acknowledge before adding new funds,[177] and the UK has banned autoplay features entirely while mandating time intervals between interactions.[178] This framework is valuable because it acknowledges that willpower alone may be insufficient against products engineered for engagement, and it creates friction that can help users close the gap between what they want and what they want to want.

Outright bans. In contexts where AI is deemed too dangerous, policymakers might prohibit its use entirely. The U.S. Space Force temporarily banned generative AI based on security concerns.[179] Several cities have banned law enforcement use of facial recognition based on privacy concerns.[180] Illinois recently banned AI therapy chatbots out of concern for user safety.[181] Recently proposed legislation in Tennessee would make it a felony to train a language model to "provide emotional support, including through open-ended conversations with a user."[182]

Advantages of Targeting This Stage

Regulating at the application layer gives lawmakers maximum context about AI-related harms. Because deployers operate in defined settings, regulators can set outcome-

---

[176] *National Strategy to Reduce Gambling Harms 2019 to 2022*, GAMBLING COMM'N (last updated Jan. 21, 2025), https://www.gamblingcommission.gov.uk/manual/national-strategy-to-reduce-gambling-harms/rts-13-time-requirements-and-reality-checks.

[177] *Id.*

[178] *Remote Gambling and Software Technical Standards: RTS 8 — Auto-Play Functionality*, GAMBLING COMM'N (last updated Jan. 21, 2025), https://www.gamblingcommission.gov.uk/manual/remote-gambling-and-software-technical-standards/rts-8-autoplay-functionality.

[179] Katrina Manson, *US Space Force Pauses Generative AI Use Based on Security Concerns*, STARS & STRIPES (Oct. 11, 2023), https://www.stripes.com/branches/space_force/2023-10-11/us-space-force-pauses-generative-ai-11671652.html.

[180] Tate Ryan-Mosley, *The Movement to Limit Facial Recognition Tech Might Finally Get a Win,* MIT TECH. REV. (July 20, 2023), https://www.technologyreview.com/2023/07/20/1076539/face-recognition-massachusetts-test-police/.

[181] Press Release, Ill. Dept. of Fin. & Pro. Reg., Gov Pritzker Signs Legislation Prohibiting AI Therapy in Illinois (Aug. 4, 2025), https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html.

[182] S.B. 1493, 114th Gen. Assemb., Reg. Sess. (Tenn. 2026); *see* also Dean W. Ball (@deanwball), X (Dec. 26, 2025, 2:28 PM), https://x.com/deanwball/status/2004635405845430607.

oriented requirements (i.e., maximum error rates for medical diagnosis or bias thresholds for hiring) and verify through real-world data that requirements are reducing harms.

This approach strongly aligns with existing regulatory expertise. Sector agencies like the FDA, NHTSA, EEOC, and SEC already police safety and integrity within their domains. Extending their mandates to cover AI-enabled products leverages existing staff, testing protocols, and enforcement tools. In many cases, new legislation may not be required because existing sector-specific laws and agency authorities can be applied to AI.

Disadvantages of Targeting This Stage

Effective application-layer intervention requires coordinating across many industries, each with its own stakeholders and lobbyists who may resist oversight. While sector-specific legislation may be more precise, policymakers may prefer omnibus "AI bills" for political reasons, since a single high-profile bill can yield greater electoral rewards than piecemeal legislation.

Even with political will for sector-specific regulation, different agencies crafting different rules creates fragmentation. AI applications that don't fit neatly into existing categories may slip through gaps. Products crossing regulatory boundaries—a health chatbot handling both insurance questions and medical advice—may face overlapping or conflicting obligations.

A further concern is regulatory capture. Sector regulators develop close relationships with the industries they oversee; over time, this can blunt enforcement or slow rule updates as technology evolves.

Summary

Intervention	Status	What It Does	Key Limitation
Incident reporting	Proposed	Rapid disclosure of malfunctions	Defining "incident"
Product liability	In effect	Sue deployers for defects	Causation difficulties
Age-appropriate design	Emerging	Protections for minors	Age verification
Anti-compulsion laws	Proposed	Session caps, break reminders, self-exclusion	Engagement trade-offs
Bans	Sector-specific	Government approval for high-risk uses	Innovation delay

The application layer is one of the most promising regulatory targets because deployers operate in concrete settings where harms are observable and measurable, and because many enforcement agencies—the FTC, EEOC, FDA, NHTSA—already have authority to regulate AI-related activities without new legislation. It is less effective when

products span multiple regulatory domains, creating gaps and inconsistencies between agencies.

The least-cost-avoider principle is useful here: deployers who control application design and the context in which AI capabilities reach users should bear responsibility for design flaws and foreseeable failures, while liability for sophisticated misuse that deployers could not reasonably anticipate or prevent should shift to end users.

Stage 6: Complementary and Enabling Platforms

Platforms are often essential for translating digital outputs into real-world consequences. E-commerce marketplaces like Amazon facilitate physical goods purchases; gig economy platforms like TaskRabbit connect requesters with workers who perform physical tasks; social media networks like Meta distribute content to mass audiences. While these platforms may not develop AI themselves, they can serve as essential channels through which AI-enabled harms materialize. A language model requires access to enabling infrastructure to purchase precursor chemicals, hire a courier, or broadcast disinformation.

This intermediary role makes enabling platforms uniquely important for AI governance. They represent the last major chokepoint before harm occurs, operating at the boundary between digital intent and physical or social consequence.

Examples of Policy Interventions

E-commerce and procurement platforms

Know-Your-Customer and suspicious activity reporting. Platforms could be required to verify purchaser identities for certain categories of goods (chemicals, laboratory equipment, dual-use materials) and report suspicious purchasing patterns to relevant authorities, similar to banking regulations under the Bank Secrecy Act.[183]

AI-aware transaction monitoring. As AI agents gain the ability to autonomously browse and purchase goods, platforms may need new detection mechanisms. Jonathan Zittrain has proposed that AI agents carry identifying credentials—akin to "license plates"—that would allow platforms to apply heightened scrutiny to agent-initiated transactions, particularly for sensitive goods categories.[184]

---

[183] Bank Secrecy Act, 12 U.S.C. §§ 1829b, 1951–1960; 31 U.S.C. §§ 5311–5314, 5316–5336.

[184] Jonathan L. Zittrain, *We Need to Control AI Agents Now*, THE ATLANTIC (July 2, 2024), https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864/.

Gig economy and labor platforms

Task screening and identity verification. Platforms like TaskRabbit, Fiverr, and Upwork could be required to maintain explicit prohibitions on tasks that facilitate illegal activity—delivering unknown substances, conducting surveillance, accessing restricted areas—and to require enhanced identity verification for task categories with higher abuse potential.

Worker protection and refusal rights. Gig workers may be unwitting participants in harmful schemes.[185] Regulations could guarantee workers the right to refuse suspicious tasks without penalty and require platforms to maintain reporting channels for workers who suspect they are being used for illicit purposes.[186]

Social media and content distribution platforms

Synthetic content labeling requirements. Building on recent state laws, policymakers could mandate that AI-generated content be labeled when distributed on social media,[187] including technical standards for embedded metadata (such as C2PA provenance standards)[188] and disclosure requirements for accounts that post primarily AI-generated content.[189]

Amplification accountability. Beyond content moderation, platforms could face obligations related to algorithmic amplification—transparency requirements for recommendation algorithms, prohibitions on amplifying content that violates platform policies, or duties to detect coordinated inauthentic behavior.

Section 230 reform. The most significant potential intervention involves modifying Section 230 of the Communications Decency Act, which currently immunizes

---

[185] *See, e.g.*, *Money Mules*, FED. BUREAU OF INVESTIGATION (last accessed Mar. 8, 2026), https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-frauds-and-scams/money-mules (explaining how criminal enterprises recruit unwitting "money mules" through false online job advertisements).

[186] *Cf.* Karnataka Platform Based Gig Workers (Social Security and Welfare) Act, 2025, Karnataka Act No. 72 of 2025, § 12(2) (India) (granting gig workers a general right to refuse tasks offered).

[187] California AI Transparency Act, ch. 291, 2024 Cal. Stat. (codified at Cal. Bus. & Prof. Code §§ 22757–22757.6) (operative Aug. 2, 2026); *id.* § 22757.3.1 (operative Jan. 1, 2027); *see also* Press Release, Consumer Reports, California Governor Signs Key Artificial Intelligence Transparency Bill Into Law (Oct. 13, 2025), https://advocacy.consumerreports.org/press_release/california-governor-signs-key-artificial-intelligence-transparency-bill-into-law/ (summarizing recent legislative developments in California).

[188] *C2PA Explainer*, COALITION FOR CONTENT PROVENANCE & AUTHENTICITY (C2PA) (last accessed Mar. 8, 2026), https://spec.c2pa.org/specifications/specifications/1.4/explainer/Explainer.html; *see also How it Works*, CONTENT AUTHENTICITY INITIATIVE (last accessed Mar. 8, 2026), https://contentauthenticity.org/how-it-works (open-source tools for C2PA compliance).

[189] Andy Duke, *How Social Media is Labeling AI-Generated Content*, KINESSO (Apr. 10, 2025), https://kinesso.co.uk/insights/how-social-media-is-labelling-ai-generated-content/.

platforms from liability for user-generated content.[190] Reform proposals range from narrow carve-outs (removing immunity for algorithmically amplified content) to broader conditions (requiring reasonable content moderation as a prerequisite for immunity).[191]

Advantages of Targeting This Stage

These platforms represent the final major chokepoint before harms materialize. Even if upstream interventions fail, enabling platforms can still prevent the final step. A bioweapon cannot be synthesized without ingredients; an influence campaign cannot succeed without distribution.

Additionally, unlike AI-specific entities, enabling platforms already operate under extensive regulatory frameworks (e.g., consumer protection laws, export controls, anti-money-laundering rules). Extending these frameworks to address AI-enabled harms leverages established compliance capabilities.

Finally, platform intermediation creates records that facilitate both ex ante screening (flagging suspicious patterns) and ex post investigation (reconstructing how harm materialized).

Disadvantages of Targeting This Stage

The sheer volume of platform activity is the first challenge. Major platforms process billions of transactions.[192] Any screening system will produce false positives and false negatives; at scale, even low error rates translate into millions of affected transactions.

Even with effective screening, platforms often cannot determine whether a transaction is AI-enabled. Without reliable signals distinguishing human from agent activity, platforms cannot easily apply differential scrutiny.[193] Requiring AI disclosure depends on the honesty of precisely those actors least likely to comply when pursuing harmful ends.

---

[190] Anna Vinals Musquera & J. Scott Babwah Brennan, *What Has Congress Been Doing on Section 230?*, LAWFARE (May 27, 2025, 3:00 PM), https://www.lawfaremedia.org/article/what-has-congress-been-doing-on-section-230.

[191] *Id.*

[192] *See, e.g.*, Community Standards Enforcement Report, META (Dec. 2025), https://transparency.meta.com/reports/community-standards-enforcement/ (describing content moderation practices for the "hundreds of billions of pieces of content produced on Facebook and Instagram in Q3 globally").

[193] *See* Alan Chan et al., *Visibility into AI Agents*, *in* FACCT '24: PROCEEDINGS OF THE 2024 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 958, 962–64 (2024), https://dl.acm.org/doi/epdf/10.1145/3630106.3658948 (proposing the use of "agent identifiers" to enable platforms to detect activity by AI agents).

Content-related interventions face an additional obstacle. For social media platforms, content-based regulations face significant constitutional scrutiny under the First Amendment. Requirements to remove or label certain categories of speech must be narrowly tailored to compelling government interests.

Finally, users seeking to evade restrictions can shift to less-regulated platforms, international services, or decentralized alternatives like cryptocurrency marketplaces or federated social networks.

Summary

Intervention	Status	What It Does	Key Limitation
KYC for sensitive goods	Partial (voluntary)	Verify purchaser identity for high-risk items	Defining scope; false positives
AI agent identification	Proposed	Require "license plates" for agent transactions	Evasion; verification difficulty
Gig task screening	Voluntary	Prohibit tasks facilitating illegal activity	Concealed purposes
Worker protection	Voluntary	Guarantee refusal rights for suspicious tasks	Detecting harmful intent
Synthetic content labeling	Emerging (state laws)	Mandate disclosure of AI-generated content	Technical feasibility; evasion
Amplification accountability	Proposed	Transparency for recommendation algorithms	Defining harmful amplification
Section 230 reform	Debated	Modify platform immunity for user content	Constitutional constraints

Enabling platforms are most valuable as intervention targets when harms require real-world resources or distribution that platforms mediate. They are the last major chokepoint where harms can be intercepted, and they benefit from existing compliance infrastructure that can be extended to AI-specific concerns. Platform-level interventions are less effective when users can easily shift to unregulated alternatives; they also face significant First Amendment constraints for content-related regulations. The least-cost-avoider principle suggests platforms should bear responsibility for harms they are well-positioned to prevent: suspicious purchasing patterns, task requests that facially violate policies, and content that clearly meets removal criteria. But platforms are not well-positioned to prevent harms requiring information they lack or judgment calls they cannot reliably make at scale.

Stage 7: End Users

End users are the final actor in the AI ecosystem. Institutional users like businesses and government agencies embed AI tools into existing workflows, often under sector-specific laws. Individual users typically lack formal compliance programs or resources to vet AI systems, yet their choices can still inflict measurable harm and are subject to liability under the tort system.[194]

Examples of Policy Interventions

Organizations using AI

Human oversight requirements. These require qualified people to oversee or override AI predictions.[195] A bank using predictive AI for loan approvals might require a human loan officer to sign off on high-stakes decisions. The idea is that human oversight provides a fail-safe to catch errors or bias that AI systems might miss, though there are compelling critiques that people might simply rubber-stamp AI decisions.[196] They are also called "human-in-the-loop" requirements.

Training and certification. Hospitals, financial firms, or law enforcement agencies might be required to obtain accredited licenses or complete mandated coursework before deploying high-risk AI systems.

Record-keeping and audit trails. Organizations may need to log prompts, outputs, and decision rationales so regulators or litigants can reconstruct how harmful decisions occurred.

Individuals using AI

Use existing tort law to find liability. Many harms individuals can inflict with AI—hacking, defamation, malpractice—are already covered by criminal statutes, tort

---

[194] Bart Custers, Henning Lahmann & Benjamyn I. Scott, *From Liability Gaps to Liability Overlaps: Shared Responsibilities and Fiduciary Duties in AI and Other Complex Technologies*, 41 COMPUT. L. & SEC. REV. 105860 (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12152026/; K.C. Halm, John D. Seiver & Edlira Kuka, *FTC Targets Tech Companies for Generative AI Fraud*, DAVIS WRIGHT TREMAINE (Feb. 22, 2024), https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2024/02/ftc-targets-tech-companies-for-generative-ai-fraud; Maarten Herbosch, *How Existing Liability Frameworks Can Handle Agentic AI Harms*, LAWFARE (Dec. 3, 2025, 10:13 AM), https://www.lawfaremedia.org/article/how-existing-liability-frameworks-can-handle-agentic-ai-harms; Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315 (2020).

[195] *See, e.g.*, Commission Regulation (EU) 2024/1689, art. 14, 2024 O.J. (L 1689), https://eur-lex.europa.eu/eli/reg/2024/1689/oj (requiring human oversight of "high-risk AI systems").

[196] *EU AI Act Shines Light on Human Oversight Needs*, IAPP (June 12, 2024), https://iapp.org/news/a/eu-ai-act-shines-light-on-human-oversight-needs.

principles, or professional ethics rules.[197] Even if the tool changes (AI), the duty does not. For example, a lawyer must still verify citations whether they come from Westlaw or a language model. State attorneys general and other entities authorized to enforce laws can reinforce this through guidance and sanctions without creating new technology-specific offenses.

Create AI-specific duties for genuine gaps. Where existing law does not squarely address new harms, legislatures can craft targeted rules—criminalizing non-consensual AI-generated intimate imagery or requiring disclosure for synthetic political ads depicting candidates saying things they never said.[198]

Advantages of Targeting This Stage

End-user liability puts accountability on the actor often best positioned to prevent harm. The individual or organization that chooses to post a deepfake, ignore a safety warning, or rely blindly on AI output controls the immediate risk and can avoid it at low cost—by double-checking facts, limiting sensitive prompts, or declining to deploy AI in high-stakes contexts.

Focusing sanctions at the user layer also protects upstream innovation. Model developers and application builders can continue releasing general-purpose tools without bearing the full weight of every possible downstream abuse, because liability attaches only when users turn tools toward prohibited or negligent uses.

Finally, user-level liability also carries expressive value. By singling out certain AI-enabled acts as punishable—non-consensual synthetic pornography, fraudulent legal filings, deceptive political ads—the law signals which uses of AI violate shared norms, helping shape social expectations in a fast-moving technological landscape.

Disadvantages of Targeting This Stage

End-user liability is inherently retroactive: it activates only after harm has occurred. For irreversible injuries—viral deepfakes that permanently taint reputations, election disinformation that cannot be "un-voted," psychological trauma from non-consensual intimate images—damages or takedowns arrive too late to help those already harmed.

---

[197] Custers, Lahmann & Scott, *supra* note 195; Halm, Seiver & Kuka, *supra* note 195; Herbosch, *supra* note 195; Selbst, *supra* note 195.

[198] *See* Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, Pub. L. No. 119-12, 139 Stat. 55 (2025) (criminalizing the publication of AI-generated non-consensual intimate imagery); *see also* Daniel I. Weiner & Lawrence Norden, *Regulating AI Deepfakes and Synthetic Media in the Political Arena*, BRENNAN CTR. FOR JUST. (Dec. 5, 2023), https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena (surveying state and federal legislative approaches to requiring disclosure of AI-generated political advertisements depicting candidates saying or doing things they never said or did).

Identifying individual bad actors is also logistically difficult. Malicious users can automate anonymous accounts, route traffic through foreign servers, or hide behind encryption, making attribution expensive or impossible. Enforcement may depend on platforms or foreign authorities with their own incentives and delays, and plaintiffs face steep proof burdens for causation, while prosecutors must establish intent beyond reasonable doubt.

First Amendment constraints present additional obstacles when regulations target individual expression. Penalties for AI-generated content must be drafted to avoid viewpoint discrimination and unconstitutional vagueness.

Finally, aggressive liability risks over-deterrence. If users fear ambiguous civil suits or criminal charges, journalists may shy away from probing models to expose bias, artists may forgo transformative remixes, and small businesses may abandon useful automation rather than gamble on uncertain legal boundaries.

Summary

Intervention	Status	What It Does	Key Limitation
Outright bans	Sector-specific	Prohibit AI use in specific contexts	May over-restrict beneficial uses
Human oversight	In effect	Require sign-off on high-stakes decisions	"Rubber stamp" risk
Training/certification	Emerging	Accreditation for high-risk AI use	Access barriers
Audit trails	Emerging	Log prompts and decisions	Storage and privacy costs
Existing law enforcement	In effect	Apply criminal/civil duties to AI conduct	May require interpretation updates
AI-specific offenses	Emerging	Criminalize deepfake porn, synthetic fraud	First Amendment constraints

End-user interventions work best when harms result from intentional misuse and the likelihood of effective deterrence is high. They are less useful when offenders are hard to identify or influence (foreign disinformation operators, anonymous bad actors). To strengthen deterrence, policymakers can offer multiple independent enforcement mechanisms (government action plus private rights of action) and impose high statutory penalties. The least-cost-avoider principle can be helpful here: where the user's choice is the proximate cause of harm, liability should follow.

Conclusion

AI policy is difficult. The technology is general-purpose, fast-moving, and embedded in an ecosystem of actors with overlapping roles and responsibilities. But

difficulty is no excuse for imprecision. This primer has argued that effective AI regulation requires specificity along three dimensions: the harm being addressed, the design principles guiding the intervention, and the stage of the AI lifecycle being targeted.

These choices inevitably involve tradeoffs. Preventive interventions reduce irreversible harms but risk being overbroad. Upstream regulations offer leverage over the entire ecosystem but are too blunt to address application-specific problems. Targeting the least-cost avoider is efficient but may concentrate compliance burdens on a small number of firms. No single intervention can address all AI harms, and most will implicate competing values.

This primer does not resolve those tradeoffs—nor could it, given how much depends on context, values, and the specific harm at issue. What it offers instead is a framework for confronting them with greater clarity. Policymakers who can specify which harm they are addressing, justify the regulatory design they have chosen, and identify where in the AI ecosystem their intervention will take effect are better positioned to write laws that are effective, proportionate, and durable.