Stanford University
Human-Centered
Artificial Intelligence

RegLab

Policy Brief
HAI Policy & Society
Stanford RegLab
November 2023

The AI Regulatory Alignment Problem

Neel Guha*, Christie M. Lawrence*[1], Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, Daniel E. Ho

WHILE THE AI ALIGNMENT PROBLEM—THE NOTION THAT MACHINE AND HUMAN VALUES MAY NOT BE ALIGNED—has arisen as an impetus for regulation, what is less recognized is that hurried calls to regulate create their own *regulatory* alignment problem, where proposals may distract, fail, or backfire.

In recent Senate testimony, OpenAI chief executive Sam Altman urged Congress to regulate AI, calling for AI safety standards, independent audits, and a new agency to issue licenses for developing advanced AI systems. His testimony echoed calls from various academics and AI researchers, who have long proposed "urgent priorities" for AI governance, including licensing procedures. Legislators have also expressed support for similar proposals. During the Altman hearing, Senator Lindsey Graham voiced support for "an agency that issues a license and can take it away." He joined Senator Elizabeth Warren in proposing an independent regulatory commission with licensing powers over dominant tech platforms including those that develop AI. Even more recently, Senators Richard Blumenthal and Josh Hawley proposed a regulatory framework featuring an independent oversight body, licensing and registration requirements for advanced or high-risk AI models, audits, and public disclosures.

1 * Equal authorship.

Key Takeaways

Although the demand for AI regulation is at a near fever pitch and may reflect a variety of legitimate concerns, four common proposals to regulate AI—mandatory **disclosure**, **registration**, **licensing**, and **auditing** regimes—are not the magic remedy to cure all that ails AI. Before rushing into regulation, policymakers should consider feasibility, trade-offs, and unintended consequences.

Many proposals suffer from what we call the "**regulatory alignment problem**," where a regulatory regime's objective or impact either fails to remediate the AI-related risk at issue (i.e., regulatory mismatch) or conflicts with other societal values and regulatory goals (i.e., value conflict).

Establishing an AI super-regulator risks creating redundant, ambiguous, or conflicting jurisdiction given the breadth of AI applications and the number of agencies with existing AI-related regulatory authorities.

Adverse event reporting and **third party-audits with government oversight** can address key impediments to effective regulation by enabling the government to learn about risks of AI models and verify industry claims without drastically increasing its capacity.

Policymakers should not expect uniform implementation of regulatory principles absent clear guidance given operationalizing high-level definitions (e.g., "dangerous capabilities") and AI principles (e.g., "fairness") is not self-evident, value-neutral or even technically feasible in some cases.

But none of these proposals is straightforward to implement. For instance, licensing regimes, at best, may be technically or institutionally infeasible—requiring a dedicated agency, as well as clear eligibility criteria and standards for pre-market evaluations—all of which would take months, if not years, to establish. At worst, a licensing scheme may undermine public safety and corporate competition by disproportionately burdening less-resourced actors—impeding useful safety research and consolidating market power among a handful of well-resourced companies. Many of these concerns are not unique to licensing, but also apply to registration, disclosure, and auditing proposals.

In "AI Regulation Has Its Own Alignment Problem," we consider the technical and institutional feasibility of four commonly proposed AI regulatory regimes—disclosure, registration, licensing, and auditing—described in the table, and conclude that each suffers from its own regulatory alignment problem.[2] Some proposals may fail to address the problems they set out to solve due to technical or institutional constraints, while others may even worsen those problems or introduce entirely new harms. Proposals that purport to address all that ails AI (e.g., by mandating transparent, fair, privacy-preserving, accurate, and explainable AI) ignore the reality that many goals cannot be jointly satisfied.

Proposals that purport to address all that ails AI (e.g., by mandating transparent, fair, privacy-preserving, accurate, and explainable AI) ignore the reality that many goals cannot be jointly satisfied.

Regulation	Description
Disclosure	Require AI system developers or deployers to share information about the system and aspects of performance, training data, design, or downstream applications with the *public*.
Registration	Require AI system developers or deployers to provide information about systems to *government regulators*, possibly accompanied by bans on use of unregistered models or penalties for unregistered use.
Licensing	Require entities or individuals to apply for and receive government approval prior to engaging in specified activities like developing or deploying AI systems, often after meeting certain criteria or demonstrating certain competencies.
Audits	Require verification by auditors that an AI system—either pre- or post-deployment—complies with relevant regulations, best practices, or standards.

2 We focus on broader proposals for regulation, noting that many specific policy proposals and recent governmental actions (e.g., the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence) include related interventions. See HAI's coverage of the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.

Without access to quality information about harms, risks, and performance, regulatory misalignment is almost assured. The current state of affairs—where only a small number of private, self-interested actors know about risks arising from AI—creates "a dangerous dynamic" between industry experts and legislators reliant on industry expertise. The question is, what policies are best situated to address the underlying problem? Rather than rushing to poorly calibrated or infeasible regulation, policymakers should first seek to enhance the government's understanding of the risks and reliability of AI systems.

Rather than rushing to poorly calibrated or infeasible regulation, policymakers should first seek to enhance the government's understanding of the risks and reliability of AI systems.

Without access to quality information about harms, risks, and performance, regulatory misalignment is almost assured.

AI Regulation's (Mis)alignment

Current calls for AI regulation are premised on addressing a wide range of dangers and market failures, from bias to environmental harms to catastrophic risk.[3] The motivation for AI regulation is to remediate these risks and market failures. However, as we demonstrate in our paper, to actually address these challenges effectively, regulatory proposals must also be tractable, well-coordinated, and *aligned* with societal values.

"Regulatory misalignment" can arise when the objectives or unintended consequences of regulatory regimes are (1) mismatched to the harm intended to remediate (what Supreme Court Justice Stephen Breyer once called "regulatory mismatch") or (2) create unacknowledged trade-offs between objectives (what we term "value conflict"). As an example of regulatory mismatch, many companies across diverse sectors employ the Equal Employment Opportunity Commission's guidance to hire protected groups at a rate that is at least 80 percent of the majority group (i.e., the "80 percent rule"). But when strictly implemented via algorithm, the 80 percent rule may resemble a quota, which is precisely what antidiscrimination law has looked at with disfavor. As an example of value conflict, consider the privacy-bias trade-off: i.e., the tension between the desire to conduct disparity assessments and informational privacy, which can make demographic data unavailable under a data minimization principle.

3 Our paper provides a more detailed list of potential harms arising from the use of AI and illustrative examples of those harms, including poor performance and inaccuracy, bias, surveillance and privacy invasion, labor displacement and job degradation, environmental costs, security, concentration of industrial power and anti-competitive behavior, geopolitical power shift, democratic erosion, and catastrophic risk.

Proposals may require technical capabilities or engineering solutions that do not currently exist.

The regulatory alignment problem also underscores the need to assess the technical and institutional feasibility of proposed AI regulatory regimes. From a technical standpoint, proposals may require technical capabilities or engineering solutions that do not currently exist. The Blumenthal-Hawley framework, for example, references watermarking requirements that may not be technically feasible yet for text-based content. Similarly, from an institutional standpoint, proposals may require government capacity that is unrealistic at best. For instance, the mere legal requirement for agencies to file AI use case inventories strained federal agencies. More ambitious proposals calling for an AI super-regulator ignore long-documented challenges of bureaucratic restructuring. Many agencies[4] already regulate AI and a new agency would either need to absorb existing authorities or coordinate with agencies, creating a recipe for turf wars. For an illustration of these challenges, look no further than the Department of Homeland Security (DHS) which many—including former DHS officials—argue simply hindered interagency coordination, rather than improving it.

Taken together, concerns of misalignment and feasibility raise two fundamental questions: (1) whether compliance with the proposed regulation will effectively address the targeted harm, without unnecessarily exacerbating other harms; and (2) whether compliance is even technically or institutionally feasible. These questions provide a starting point for any sensible regulatory scheme.

Challenges to Achieving AI Regulatory Alignment

The regulatory proposals described share common challenges. First, from a technical perspective, defining the scope of AI regulation—what AI systems and which system updates are subject to regulations—is exceedingly difficult. One common proposal is to define "frontier" models by amount of compute expended, model size, or model performance, like achieving at least a 1300 on the SAT. But regulations based on threshold criteria may create incentives for strategic evasion (e.g., developing multiple models below the compute threshold and combining their outputs) or may fail to address risks stemming from smaller but nonetheless powerful models (e.g., thresholds based purely on size or compute)—raising questions as to whether proxy criteria provide a meaningful representation of risks and challenging motivations for confining the bulk of regulation to the "frontier."

Second, from an institutional perspective, agencies have a limited supply of technical talent to lend the expertise necessary for compliance or enforcement. With fewer than 1 percent of new AI Ph.D.'s choosing to forego lucrative private sector salaries in favor of

4 For example, the Food and Drug Administration regulates AI medical devices, the Department of Housing and Urban Development conducts oversight over algorithmic bias in housing, the Consumer Financial Protection Bureau regulates AI used in consumer financial products, the Consumer Product Safety Commission ensures safety in consumer products, the Federal Trade Commission regulates advertising claims and enforces consumer protections, the Department of Transport has oversight of over self-driving cars, the Equal Employment Opportunity Commission examines AI used in employment decisions, and the Securities and Exchange Commission has rulemaking around the use of AI by broker-dealers or investment advisors, to name a few.

government employment, the shortage of technical experts in government poses a profound challenge for regulatory design and implementation. When Eric Schmidt, chair of the National Security Commission on AI (NSCAI), was asked during a House oversight hearing in March about the implementation of NSCAI recommendations, he singled out one acute need: technical talent in government. You can't regulate what you don't understand.

Third, proposed regulations may give the appearance of a solution but, in fact, be hollow. For instance, Kai-Fu Lee, a prominent investor and former President of Google China, refers to an emerging consensus around audits to create market pressure for responsible AI, likening them to the emerging audit ecosystem for environmental, social, and governance (ESG) factors. But this analogy may be the wrong one: With malleable standards and auditors paid as hired guns, audits may simply act as rubber stamps. And a number of critical risks of AI may be better addressed through conventional regulation, rather than AI-specific policies. For example, there is much anxiety about how open foundation models may decrease the barriers preventing bad actors from creating bioweapons. But safeguards intended to increase the safety alignment of more closed foundation models can be easily stripped away. Even if usage can be monitored with closed models, some concerns about safety and catastrophic risk may be more readily addressed by tightening the regulation of laboratories that provide the infrastructure for bioweapon production.

A number of critical risks of AI may be better addressed through conventional regulation, rather than AI-specific policies.

We cannot navigate trade-offs when blind to them.

Fourth, policymakers will inevitably encounter trade-offs between societal values and regulatory objectives—no single proposal can achieve every regulatory goal. Proposals commonly call for AI regulation to promote systems that are maximally accurate, safe, effective, nondiscriminatory, privacy-preserving, and transparent, but the reality is that these goals cannot always, or in some contexts ever, be jointly satisfied. Data minimization, for instance, has meant that federal agencies have lacked demographic data to carry out legally mandated equity assessments, posing a privacy-bias trade-off. We cannot navigate trade-offs when blind to them.

Last, reliance on insights from industry potentially creates opportunities for capture. The number of entities lobbying on AI-related issues increased from 30 in 2017 to 158 in 2022. Several companies have voiced support for forms of licensing, arguing that it will ensure more responsible use and development, yet these advocates may also be the beneficiaries of such regulation. In other domains, estimates suggest occupational licensing can reduce employment and raise prices without necessarily improving the quality of goods or services. Put differently, the motivation for licensing may be the risk of growing competition.

Policy Recommendations

Our analysis supports four concrete recommendations.

First, adverse event reporting—both mandatory and voluntary—can address a central impediment to effective AI regulation—the lack of reliable information about different AI systems and their risks. For instance, the government could require the transparent reporting of deleterious AI behavior, ranging from concrete harms (e.g., misdiagnosis by medical AI systems) to more abstract concerns (e.g., generation of biological pathogens). The agency managing the reporting system could then either refer incidents to existing agencies, or identify gaps if reports fall between existing authorities. Such a policy would require relatively few technical and institutional resources to operationalize and provide clear benefits. Previous experience with incident reporting systems (e.g., the Food and Drug Administration's Adverse Event Reporting System and Cybersecurity and Infrastructure Security Agency reporting system) has shown the value of incident reporting for identifying threats and informing future regulation. Lightweight registration of models can be seen as a complement to adverse event reporting, enabling regulators to understand what models might be susceptible to similar risks.

Second, government oversight of third-party auditors can enable the verification of industry claims without miring the government in direct auditing of AI systems. Third-party audits would ensure independent assessment of AI algorithms and models and verification of industry claims. Reducing conflicts of interest will be critical. Government oversight—similar to the Public Company Accounting Oversight Board—may encourage standardization in the third-party audit market and improve the quality of information available about AI systems and their performance.

Proposals that strengthen authorities of existing agencies, including the many agencies already regulating AI, are more likely to successfully implement timely regulation than proposals reliant upon a new super-regulatory agency.

Third, proposals that strengthen authorities of existing agencies, including the many agencies already regulating AI, are more likely to successfully implement timely regulation than proposals reliant on a new super-regulatory agency.

Fourth, policymakers must grapple with the reality that regulatory regimes expose tensions between objectives (e.g., restrict foundation models versus democratize AI innovation) and values (e.g., fairness versus privacy) that cannot be avoided simply by demanding the technical community operationalize yet-to-be-determined standards and metrics.

In sum, there is a role for reasonable government action to govern AI. But AI's regulatory alignment problem is a hard one. It will only be made harder if policymakers rush to regulate without regard for the feasibility or unintended consequences of their proposals.

Reference: The original article is accessible at Neel Guha, Christie M. Lawrence et al., "**AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing**," *George Washington Law Review, Symposium on Legally Disruptive Emerging Technologies* (Forthcoming), https://dho.stanford.edu/wp-content/uploads/AI_Regulation.pdf.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.

Stanford University's Regulation, Evaluation, and Governance Lab (RegLab) partners with government agencies to design and evaluate programs, policies, and technologies that modernize government.

The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**


Neel Guha is a JD/PhD student in computer science at Stanford University.


Christie M. Lawrence is a concurrent JD/MPP student at Stanford Law School and the Harvard Kennedy School.


Lindsey A. Gailmard is a postdoctoral scholar at the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford University.


Kit T. Rodolfa is the research director at the RegLab at Stanford University.


Faiz Surani is a research fellow at the RegLab at Stanford University.


Rishi Bommasani is a PhD student in computer science at Stanford University and the society lead at the Stanford Center for Research on Foundation Models (CRFM).


Inioluwa Deborah Raji is a PhD student in computer science at UC Berkeley.


Mariano-Florentino Cuéllar is the president of the Carnegie Endowment for International Peace, a former Supreme Court Justice of California, and the former Stanley Morrison Professor of Law at Stanford.


Colleen Honigsberg is Professor of Law, Bernard Bergreen Faculty Scholar, Faculty Co-Director of the Arthur and Toni Rembe Rock Center for Corporate Governance, and Senior Fellow at the Stanford Institute for Economic and Policy Research (SIEPR).


Percy Liang is Associate Professor of Computer Science, Senior Fellow at HAI, and Director of CRFM.


Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science and Computer Science (by courtesy), Senior Fellow at HAI, Senior Fellow at SIEPR, and Director at the RegLab.

Stanford University
Human-Centered
Artificial Intelligence

RegLab

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**