Christoph Winter & Charlie Bullock

Radical Optionality

Governing Transformative AI Under Uncertainty

INSTITUTE
FOR LAW & AI

CONTENTS

Christoph Winter is Assistant Professor of Law and AI at the University of Cambridge.
Both authors are affiliated with the Institute for Law & AI.

Introduction

The prospect of "transformative AI" appears to present policymakers with a dilemma: overregulation could stifle innovation and forfeit the potential benefits of the technology, while a failure to regulate appropriately could have disastrous implications for public safety and national security.

It's true that security and innovation are sometimes in tension. Some safety measures do impose costs on innovation, and some forms of deregulation do carry genuine risks. But there is also a class of policies that would meaningfully increase safety without imposing significant costs on innovation. We argue that governments should aggressively implement these policies; this is the main thrust of the governance strategy discussed in this essay, which we call "radical optionality."

At its core, radical optionality is about preserving democratic governments' ability to make good decisions about how to govern transformative AI systems as circumstances evolve. In the short term, this means avoiding overregulation while rapidly building the institutions, information channels and legal authorities needed to respond competently to a broad range of scenarios.

The argument for focusing on optionality is simple, and—if you accept a few reasonable assumptions—compelling. These assumptions are:

1. That there is a real possibility of transformative AI (defined as "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution") being developed within the next ten years;
2. That profound uncertainty exists as to what capabilities transformative AI systems will possess, what benefits and risks they will generate, and what the best ways for society to capture the benefits while mitigating the risks will be;
3. That a transformatively impactful dual-use technology with significant national security implications will inevitably require some degree of government oversight; and

4. That building the institutional capacity required to effectively govern transformative AI systems will take years, and that society therefore cannot afford to wait until transformative capabilities have actually been developed.

Justifying the first assumption is beyond the scope of this paper. Whether "AGI" or "superintelligence" or "powerful AI" or "transformative AI" will ever arrive, and when, are questions that have been debated extensively elsewhere. But if you believe that transformative AI is possible, we hope to demonstrate that the case for radical efforts to preserve optionality is overwhelmingly strong.

The Challenge of Regulating Transformative AI

The importance of getting the regulatory response to a truly transformative technology right is obvious. The complexity of the problem may not be. Scholars who study the regulation of emerging technologies have long acknowledged the difficulties posed by the "pacing problem." In brief, technological progress often occurs at such a rapid pace that laws, regulations, and the legal system are unable to adapt quickly enough to keep up. This makes it difficult for policymakers to effectively govern emerging technologies. In the AI context, this problem is compounded by the fact that AI systems are in some ways uniquely difficult to understand, and may possess capabilities that even the people who created them are initially unaware of. Capability profiles are also jagged: a model might perform at expert level on one task while failing at much simpler tasks in another domain, making it hard to infer a system's overall competence from any given benchmark. In other words, AI governance involves decision making under extreme uncertainty about the future capabilities of the technology, the nature and severity of its risks it might pose, and the benefits it might offer.

By some measures, AI model performance has been improving exponentially, and some researchers believe that this trend will continue in the coming years. One particularly exciting and concerning prospect is the possibility of

recursive self-improvement. As increasingly capable AI systems are developed, perhaps with superhuman programming and/or research abilities, these systems might facilitate the development of even more capable systems, which might facilitate the development of still more capable systems, and so on. An early version of this phenomenon is already occurring: engineers at some frontier labs have reported that a significant majority of the code used to create next-generation AI systems is being written by current-generation AI systems. And because AI systems write code far more quickly than human programmers, recursive self-improvement may result in dramatic, and perhaps exponential, improvements in capabilities in the relatively near future.

Because human beings are psychologically disinclined to accept the implications of exponential growth, the exponential progress of AI capabilities research further compounds the difficulty of governing transformative AI. Historically, institutions have often failed to grapple with the reality of exponential trends until it was too late to respond effectively. For instance, epidemiologists in early 2020 dismissed COVID-19 as being less prevalent than the flu, and the International Energy Agency systematically underestimated solar power growth for over a decade, repeatedly predicting that it would level off or decrease when in fact the industry maintained roughly 25% annual growth.

All this is to say that we have every reason to believe that regulating AI effectively will be unusually difficult even in comparison to past efforts to regulate emerging technologies, which have been far from universally successful.

Approaches to Regulation

Let the Market Handle It

In light of these challenges, how should the government regulate transformative AI? One possible answer is that it shouldn't. Libertarian writers like Adam Thierer have made the case for a culture of "permissionless innovation" for AI development, in which the role of government would be limited to enforcing existing laws and facilitating industry self-regulation with "soft law" tools such as voluntary standard-setting. In Europe, Mario Draghi's competitiveness review strikes a similar note, urging the EU to "profoundly refocus its collective efforts on closing the innovation gap with the US and China, especially in advanced technologies."

The appeal that this view has for techno-optimist champions of innovation isn't hard to understand. Historically, governments have often struggled to regulate emerging technologies in a way that does more good than harm. If you squint, calls for regulation to preemptively mitigate speculative future AI risks look a lot like historical calls to (over)regulate nuclear energy, which arguably led to disastrous economic and environmental consequences in the form of missed opportunities to generate cheap and abundant clean energy. Hundreds of thousands of lives have likely been lost prematurely due to air pollution that could have been prevented by removing the regulatory barriers that prevent nuclear energy from being cost-effective. And the respective trajectories of the tech industries of Europe and the United States over the course of the last few decades is like something out of an Ayn Rand novel; it gives rise to the same instinctive sense of bewildered contempt in the breast of the libertarian observer as a satellite photograph of the Korean peninsula at night. It's easy to see why someone with these beliefs would be disinclined to follow the lead of those who have succeeded only in regulating their own tech industry out of existence,[1] and more disposed to side with builders and visionaries than takers and bureaucrats.

But, ironically, support for this techno-optimist perspective typically depends (at least tacitly) on skepticism about the future trajectory of AI capabilities. Again, consider the example of nuclear power. However well libertarian objections to the overregulation of nuclear power plants have aged, it's much harder to make the case that the advent of the nuclear age

required no new laws or regulations whatsoever. Presumably, reasonable people can agree that a laissez-faire approach to regulating the private acquisition and possession of nuclear weapons would be inadvisable. “Every private individual should be allowed to buy as many nuclear weapons as they want, free from government interference” is a very difficult view to defend with a straight face, no matter how strong your pro-market priors.

A truly transformative general-purpose AI system would likely have significant military applications, perhaps even as significant as the military applications of nuclear fission. Many national security commentators predict that the importance of AI on the battlefield will continue to rapidly increase. If this is the case, we should not expect a totally laissez-faire approach to AI governance to be any more practically or politically feasible than a laissez-faire approach to the governance of nuclear weapons. In other words, a truly transformative dual-use[2] technology will almost certainly require a nonzero amount of regulation.

A crucial difference between nuclear weapons and transformative AI is that nuclear weapons unquestionably exist, while transformative AI is still only a possibility. Promulgating rigid and detailed regulations addressing nuclear weapons before they were even being developed would likely have been foolish, given the well-documented difficulties that regulators have historically had in predicting the future course of technological progress. But at some point, once a transformative dual-use technology is actually under development, regulation becomes unquestionably necessary—and at that point, all stakeholders have a mutual interest in ensuring that regulations are competently designed, interpreted and enforced. Insufficient government regulatory capacity could lead to hamfisted and overly harsh regulation down the line, once stakeholders realize that society is about to be fundamentally transformed. Worse yet, an unprepared government might regulate incompetently, harming industry without helping the public.

If that’s the case, then, given the stakes, wouldn’t it be a good idea to start preparing? Even if “let the market handle it” is typically sound wisdom, that heuristic alone can’t fully resolve issues like this one where the national security implications of a particular technology are likely to necessitate some degree of government oversight at some point. And given the scale of the costs and benefits at issue, “let the market handle it until these national security issues manifest themselves, if they ever do” isn’t an adequate

solution either. If there are steps that can be taken to increase the regulatory capacity of the relevant government bodies *without* significantly inhibiting innovation, it's incredibly important that we should take them as soon and as well as possible.

Anticipatory Governance and the Precautionary Principle

On the other side of the aisle/pond, some will object to radical optionality from the opposite direction, arguing that we need to implement strict prescriptive regulation (such as a law prohibiting AI research, or an FDA-style licensing regime that would only authorize companies to develop systems already proven to be safe) as soon as possible. These critics sometimes suggest that the regulatory response to the possibility of transformative AI should observe the precautionary principle, or that regulators should try to predict the future of AI progress and implement an anticipatory governance approach.

In its most extreme form, the precautionary principle dictates that any action which might pose a risk to public health or safety should be prohibited unless the party wishing to undertake the action can prove that the action is not dangerous. Because every truly revolutionary technology creates risks as well as benefits, this "hard" precautionary principle would prohibit development of transformative AI for the foreseeable future. This version of the precautionary principle is simply bad policy, because it ignores the possibility that regulation might itself cause more harm in expectation than the risks that the regulation is intended to address. However, there are also a number of less unreasonable alternative formulations. The EU, for example, endorses a version of the precautionary principle that involves conducting a cost-benefit analysis that takes both the costs of regulating and the costs of failing to regulate into account; this is the precautionary principle that the EU's Code of Practice for General-Purpose AI invokes. And some scholars have argued that the precautionary principle is justified in cases where there is a real danger of truly catastrophic harm, because even a very low probability of a cost that can be said to be "infinite" (such as the extinction of all life on the planet) outweighs even a very high probability of a very substantial finite benefit.

A precautionary principle that incorporates cost-benefit analysis might not be inconsistent with radical optionality, under reasonable assumptions about the costs of preemptively regulating emerging technologies. And while the catastrophic harm argument is theoretically sound, it doesn't have obvious implications for AI governance. For one thing, it isn't clear that restricting innovation in liberal democracies would lower the overall long-term probability of catastrophic outcomes, given that AI research would almost certainly continue to be conducted in authoritarian states abroad. For another, we're not convinced that the potential risks associated with transformative AI are necessarily more infinite than the potential benefits. To take one obvious example, powerful AI systems in the right hands could prevent a catastrophe that would otherwise have occurred, such as a nuclear war or a global pandemic. More generally, it's plausible that a wealthier society with more access to intelligence will be more willing and able to invest in averting catastrophe, meaning that growth and progress may in fact be anti-correlated with catastrophic risk.

There's no scientifically certain way of determining, from our current vantage point, whether the potential benefits of transformative AI outweigh the potential risks. In the face of this kind of uncertainty, relying on rough heuristics may be the best we can do. The precautionary principle is an application of one possible heuristic—namely, that policy should treat changes as likely to be harmful unless they can be proven harmless. The almost equally rough heuristic that we prefer is based on the observation that throughout history new technologies have typically (though not invariably) produced a net benefit for society in the long term.[3]

Unlike most precautionary principle approaches, anticipatory governance doesn't rely on techno-pessimist assumptions. Instead, proponents of anticipatory governance are optimistic about the ability of regulators to predict the course of technological progress and implement effective regulations before they're needed. The great advantage of this approach is that it can prevent harms before they happen. It's not unreasonable to expect an ounce of prevention to be worth a pound or more of cure.

Our objection to an anticipatory governance approach is based on the observation that predicting the future trajectory of technological progress is difficult, and that governments have historically been terrible at it. We've written about this problem in more detail elsewhere, but essentially, we think

that the history of regulation of emerging technologies shows that attempts to address risks that are poorly understood often result in legal regimes that are ineffective or even counterproductive. The hands-off approach that the U.S. government took when the Internet was coming into existence, for example, holds up better in hindsight than the same government's attempts to anticipatorily regulate home taping via the Audio Home Recording Act of 1992, which was rendered mostly obsolete almost as soon as it was passed by the advent of the personal computer.

An obvious corollary of the fact that it's very hard to predict the future course of technological development is that it's also very hard to predict how society will interact with unknowable future technologies and what regulatory interventions will be necessary or advisable. Also, regulatory regimes tend to be much stickier and more path-dependent than you might expect, so changing course after initially committing to a certain regulatory approach may be difficult and costly. Crucially, institutional lead times are long; recruiting scarce expertise, establishing secure testing and information-sharing, and designing flexible, lawful authorities takes years. If capacity-building begins only once risks and benefits are unmistakable, the decisive window for proportionate action will likely have closed.

Some degree of anticipation is necessary, of course. Even our suggestion that governments should focus on building capacity is based on predictions—we predict that transformative AI systems may be developed and that, if developed, they will likely be dual-use. Decision making under uncertainty is unavoidable, and radical optionality is a formula for minimizing uncertainty, not for eliminating it. But we think there's a great deal to be gained by minimizing uncertainty and making informed decisions.

If you find yourself unconvinced by the points we raise above, and still favor a more restrictive regulatory approach, fair enough. Reasonable people can disagree. But the final argument in favor of an optionality-focused approach is simpler and harder to dispute: the policies we favor are more politically realistic, at least in the short term. California and New York have passed transparency laws that require frontier AI developers to publish their policies for addressing catastrophic risks, and the EU's General-Purpose AI (GPAI) Code of Practice includes similar transparency and reporting requirements. But far more sweeping and costly policies, like an international treaty prohibiting frontier AI development or a moratorium on data center

construction, have no realistic chance of becoming law in the near future, either in the U.S. or Europe. To be clear, we have our doubts about whether such proposals would be desirable or even workable—but if you disagree, you should still support the policies we're recommending even as you continue to advocate for more ambitious proposals.

Radical Optionality

Instead of regulating or failing to regulate, governments can *prepare* to regulate in a way that will improve their ability to respond to a wide range of possible scenarios, foreseen or unforeseen. This can be done by building strong regulatory institutions, equipping them with appropriately flexible authorities, and ensuring that they have access to the information they'll need to respond competently and decisively if it ever becomes clear that regulation is needed to address some intolerable risk to public safety or national security. Unlike the precautionary principle approach, the policy measures involved would impose only negligible burdens on AI companies and would have a negligible impact on innovation. Instead, the costs of an optionality-preserving approach would be measurable in taxpayer dollars and political capital.

This is where the "radical" comes in. Simply focusing on preserving optionality sounds like a rather moderate proposal. But we argue that governments should be willing to spend an *extraordinary* amount of money, effort, and political capital on preserving optionality.

This follows from the scale of the problem—again, the premise from which we're starting is that there is a distinct possibility of a transition at least as significant as the Industrial Revolution occurring over the course of the next few years or decades. If governments genuinely accept that dual-use transformative AI systems may arrive in the near- or medium-term future, the logical consequence is that a great deal of what they can do to even marginally improve the odds of the transition going well will be cost-justified. The primary costs at issue, then, are costs to innovation. Governments should be wary of *counterproductive* interventions, but not much concerned with the actual pecuniary cost of any realistic measure that seems likely to have net-positive results. Even if there's a 95% chance that the money spent on a given policy measure is wasted, a five percent chance of some positive impact

in terms of mitigating the risks or realizing the benefits of the most important invention in human history would mean that the costs were justified a thousand times over in expectation.

Radical optionality should be distinguished from "muddling through," the incrementalist policy making philosophy associated with Charles Lindblom. Both approaches are skeptical of committing to grand regulatory regimes based on present-day guesses about how the future will unfold, and both prefer flexibility to foreclosure. But muddling through tends to preserve flexibility largely by default, through inaction and delay. Radical optionality, by contrast, requires proactive investment in building institutional capacity—not enough to lock in a particular regulatory future, but enough to make good choices when they become necessary.

Of course, there may come a time when preparation ceases either to be necessary or to be sufficient. If progress in AI capabilities research plateaus and it becomes clear that transformative systems are *not* on the horizon, the policies we're suggesting would mostly be redundant. On the other hand, if transformative AI is imminent and dire risks to public safety are manifesting, a substantive regulatory response could be both necessary and politically inevitable.

Consider former OpenAI employee Leopold Aschenbrenner's series of essays on "Situational Awareness." Aschenbrenner proposes that it is necessary and "inevitable" for the U.S. government, motivated by national security concerns, to prohibit private companies from working on transformative AI systems and instead invest trillions of dollars in a government-run "AGI Manhattan Project," starting in perhaps 2027 or 2028. Unsurprisingly, this take generated a fair amount of criticism as well as its share of approbation. In another influential recent piece, the crypto pioneer Vitalik Buterin proposed a way of *avoiding* a future in which transformative AI is developed in a closed, centralized, securitized context. His solution was a philosophy that he calls defensive acceleration, or "d/acc," which would focus on the development of defense-favoring technologies.

These two perspectives are profoundly different, but they both generally deal with the problem of balancing the need to encourage innovation and the need to mitigate risks. Radical optionality is a strategy for doing exactly this—or rather, a strategy that takes advantage of the fact that safety and innovation

aren't necessarily conflicting values. The appeal of this approach is that it promises to work well in a wide variety of futures. Regardless of whether you agree with Aschenbrenner or Buterin (or, as will be the case for most people, with neither) about the right way to handle the development of transformative AI when it happens, the correct course of action at this point in time is to avoid overregulating for now while preparing our institutions for the challenges that may lie ahead.

In the scenario Aschenbrenner predicts, the U.S. government will eventually and suddenly realize the necessity of urgent action and, like a student pulling an all-nighter before an exam, throw together the most complex public project in human history at the last minute. We don't necessarily agree with Aschenbrenner's timelines or his certainty about how transformative AI systems will be developed. But if he does turn out to be correct, the project he contemplates will produce better outcomes if the government is better prepared to spring into action—if it has access to useful information, qualified personnel, and flexible governance mechanisms. And if, like Buterin, you want to avoid a future where transformative AI's development is centralized and securitized, an open, democratic future is more likely to come about if governments have good options available to them other than springing into action to avoid imminent disaster.

Concrete Policy Suggestions for Preserving Optionality

All this isn't simply a long-winded way of saying that we should wait and see. There are a number of concrete first steps that can and should be taken as soon as possible to increase the regulatory capacity of governments in the U.S. and Europe without creating any significant barriers to innovation. These policies would create meaningful optionality, ensuring that we have the information, expertise, and institutional capacity needed to make effective choices when critical decision points arrive rather than being forced into suboptimal responses by a lack of preparation.

Information-Gathering Authorities: Reporting and Transparency Requirements

One top priority for preserving optionality is the implementation of well-designed information-gathering authorities. Like corporations, LLMs, and other complex systems, government agencies thrive on a diet of information; it's been said that "information is the lifeblood of good governance." It's a generally accepted fact, supported by reams of legal scholarship as well as by basic common sense, that government agencies make better decisions when they have access to better information. By beginning to gather information about frontier models now, governments can develop expertise in securely collecting, analyzing, and sharing information about advanced AI systems. On the other hand, if information-gathering doesn't begin in earnest until the advent of transformative AI systems makes it undeniably necessary, the agencies charged with collecting and processing information will lack institutional expertise and make worse decisions. This would impose costs, not just on the public, but also on AI companies whose sensitive and valuable information would be processed less securely and efficiently.

Information-gathering authorities can be grouped into two broad categories: transparency requirements, which mandate that companies publish certain information about their models publicly, and reporting requirements, which require companies to share information with a government agency. Transparency requirements have the advantage of allowing academics, civil society organizations, independent researchers, and industry groups to review the disclosed material, increasing the total resources that can be brought to bear on analyzing disclosures. But requiring companies to reveal trade secrets or other sensitive business information publicly carries costs to innovation. The advantage of reporting requirements (a category that includes incident reporting requirements) is that more detailed and sensitive information can be collected without compromising public safety or company trade secrets. Both transparency and reporting requirements help preserve optionality by improving governments' ability to competently assess if, when, and how to regulate frontier AI systems.

Most significant AI safety legislation enacted to date consists primarily of either reporting or transparency requirements. The EU's GPAI Code of

Practice contains a number of reporting requirements. And the most significant state AI laws enacted in the U.S., such as California's Transparency in Frontier Artificial Intelligence Act (SB 53) and New York's RAISE Act, consist primarily of transparency requirements. This is a sensible approach, because any more ambitious substantive regulation introduced in the future will benefit from being designed and implemented by better-informed actors, and information-gathering authorities are typically minimally burdensome and easy to enforce.

Comparing the GPAI Code of Practice's reporting requirements with state transparency proposals illustrates the respective contributions of transparency and reporting requirements. The GPAI Code of Practice requires frontier AI companies to create (and sometimes, but not always, send to a regulator, the European AI Office) a detailed "safety and security framework" documenting information relating to the company's risk mitigation efforts and a separate "safety and security model report" describing how a given model complies with the framework. RAISE, SB 53, and similar state transparency laws, by contrast, require companies to publish a "safety and security protocol" or some equivalent document, essentially making it mandatory for frontier AI companies to produce documents similar to the policies and frameworks that they have already been producing on a voluntary basis. The exact requirements for what this framework must contain vary, but in general the information state transparency laws require is, appropriately, quite limited in comparison to what the Code of Practice requires.

Transparency and reporting requirements are complementary. Governments can increase optionality by using reporting requirements to ensure that risk-relevant information is collected by government offices that can be trusted to securely process it, while using transparency requirements to facilitate public access to non-sensitive information that companies can produce without incurring competitive harm. This benefits the public because it allows civil society to bring specialized knowledge and resources to bear on analysis of public disclosures while empowering the appropriate government body to concentrate its resources on analyzing (and potentially responding to) any concerning information in the nonpublic reports.

State governments generally lack the institutional capacity to securely process and make use of the more detailed information that reporting

requirements are well-suited to producing. Also, the kind of information that reporting requirements produces would generally be of limited value to individual state governments. Therefore, while state transparency requirements are a decent substitute for a federal transparency framework (as long as the state requirements are harmonized so as to avoid a patchwork that would impose significant compliance costs on companies), reporting requirements are better implemented at the federal level. Previously, the U.S. federal government collected information about frontier models via reporting requirements administered by the Bureau of Industry and Security, pursuant to a proposed rule issued during the Biden administration, which required companies building the most advanced systems to report information like the results of internal safety evaluations and the physical and cybersecurity measures taken to protect model weights.

Notably, the fundamental idea of this kind of information-gathering didn't meet with any objection from the AI companies that would have been subject to the rule during the notice and comment phase of the rulemaking process, although companies like OpenAI and Anthropic suggested changes to the nature and frequency of the requirements and emphasized the importance of securely handling the reported information. Some conservatives objected to the legal basis for the reporting requirements because the Biden administration relied on Defense Production Act (DPA) authorities in promulgating them, and the proposed rule was eventually abandoned after President Trump revoked the Biden-era executive order on AI.

But setting aside the issue of whether invoking the DPA to authorize reporting requirements is appropriate or legal, reporting requirements are obviously a good idea as a matter of policy. They are extremely light touch—the total burden of the proposed BIS rule would essentially have been that approximately five companies each had to send one email to BIS once every few months. Also, as discussed above, they can be invaluable for informing agency decision making. It seems unlikely that the Trump administration will replace the defunct BIS reporting requirements with new information-gathering authorities, but for the record, doing so would be a good way to operationalize the White House AI Action Plan's stated goals relating to information-sharing and evaluating national security risks in frontier models.

Once solid transparency and reporting requirements have been established, the next step is to establish an auditing regime so that independent third-party or government auditors can verify (a) the adequacy of a company's policies for addressing risks, and (b) the company's compliance with their own plan. Auditing requirements are not quite as light-touch as transparency or reporting requirements, but the burden to companies is still fairly minimal as long as the requirements are well-designed and issues like secure handling of information and protection of trade secrets are adequately addressed. Promisingly, some leading AI companies have affirmatively requested auditing requirements.

Whistleblower Protections

Whistleblower protections are another important tool for increasing government access to information about frontier AI systems. Ideally, whistleblower protections would ensure that employees at frontier AI companies could report information about risks posed by frontier AI systems, without fear of retaliation, to an appropriate government office that has expertise in handling sensitive information securely. Like transparency and reporting requirements, this would increase the government's access, in expectation, to important information about the risks posed by advanced AI systems. And like those requirements, well-designed whistleblower protections do not burden innovation to any significant degree, as they impose virtually no positive obligations on affected companies. This is not to say that whistleblower protections are totally costless; they could lead, for instance, to a marginal increase in the cost of defending against frivolous whistleblower lawsuits filed by disgruntled former employees. But this is not the sort of cost that seriously hinders innovation, and at the end of the day "don't fire or punish an employee for telling the government about a serious, specific risk that your research creates" is a very reasonable ask.

Currently, most frontier AI company employees are entitled to the whistleblower protections of California law, which protects employees from retaliation for reporting violations of any state, federal, or local law or regulation. California's new frontier AI transparency law, SB 53, adds to these generally applicable protections by protecting certain frontier AI company employees from retaliation for disclosures about catastrophic risks. But SB 53's protections are quite limited, applying only to employees "responsible

for assessing, managing, or addressing" catastrophic risks. There's still a need for additional federal whistleblower legislation to universalize these protections and to protect the secure disclosure, to a designated federal agency, of information about significant risks to public safety or national security even when no law has been broken. A number of existing U.S. federal statutes recognize the need for this kind of protection in other industries—for example, the Federal Railroad Safety Act protects the reporting of any "hazardous safety or security condition" related to railroads. Because AI is an emerging and largely unregulated technology, it's easy to imagine situations in which a given model's capabilities might pose a real danger to public safety even in the absence of any clear violation of law by its developer. Recent reporting about the cyberdefense capabilities of Anthropic's Claude Mythos Preview, which has identified "thousands of zero-day vulnerabilities, many of them critical," illustrates this point; releasing a similarly capable model publicly without warning would likely not have been illegal, but might nevertheless have caused significant harm.

The AI Whistleblower Protection Act, a bipartisan bill introduced by Senator Chuck Grassley of Iowa, would fill this gap in existing whistleblower protections by prohibiting retaliation against whistleblowers who disclose information about "substantial and specific" dangers to public health, public safety, or national security to an appropriate government agency. Passing that bill, or something like it, would be a solid first step towards building the kind of U.S. government capacity for securely gathering and processing information about risks from frontier AI systems that is needed to preserve optionality.

In the EU, whistleblower disclosures about violations of EU law are entitled to some protection. Starting in August 2026, this protection will extend to disclosures about violations of the EU AI Act. Disclosures about dangers to public safety that don't involve EU law violations, however, are generally not protected. Because most frontier AI companies are headquartered in the U.S., Europe's ability to collect information about risks from the most advanced models will likely be significantly more limited than that of the U.S. Even so, it would be wise for the EU to implement something like the AI Whistleblower Protection Act for companies operating in the EU, to preserve optionality in the event that dangerous capabilities are developed there in the future.

Information-Sharing

Securely gathering and sharing information within and between governments, and between governments and outside stakeholders, is an important priority as well. The government's role as a coordinator and facilitator of discussions between a variety of stakeholders is difficult to entirely replace with private governance mechanisms. Deregulatory measures can do some of the work here, in the form of antitrust safe harbors or guidance clarifying that labs sharing specific sanctioned kinds of safety-relevant information with each other through the Frontier Model Forum or similar bodies doesn't violate antitrust laws.

It might appear that the U.S. has nothing to gain from international information-sharing, given that AI innovation to date has primarily taken place in the U.S. and given that the U.S. seems poised to reap the lion's share of the rewards. But in the long run, the U.S. stands to gain from sharing and receiving some information about model capabilities with and from close allies, as when the UK AI Safety Institute shares the results of its pre-deployment evaluations of frontier models with its U.S. counterpart, the Center for AI Standards and Innovation (CAISI), or when the two agencies conduct joint pre-deployment evaluations together. Wholesale objections to any degree of international cooperation simply don't make sense unless they're based in skepticism about the possibility of transformative AI; at some point, a technology that reshapes the global economy will inevitably require some degree of coordination between allies. The EU AI Act provides a promising authority for intergovernmental information-sharing at Article 78(5), which authorizes the European Commission and EU Member States to exchange "confidential information with regulatory authorities of third countries with which they have concluded bilateral or multilateral confidentiality arrangements guaranteeing an adequate level of confidentiality."

Even "Situational Awareness," which posits not only the likelihood but the absolute *certainty* of superintelligence being developed exclusively by and for the U.S. national security enterprise, recognizes the importance of a "tighter alliance of democracies" for pooling resources and protecting supply chains. Anthropic CEO Dario Amodei's essay "Machines of Loving Grace," which has also been criticized for its focus on establishing western, and specifically American, AI dominance, suggests an "entente strategy" for bringing together

a “coalition of democracies” to exercise control over the AI supply chain and distribute the benefits of powerful AI in order to promote democracy. In these and every other sensible proposal for internationally coordinating AI governance efforts, the first step is to establish channels for securely sharing appropriate information about model capabilities and risks. This sharing wouldn’t impose any significant regulatory burden on labs, but it would prime the pump for further cooperation when and if it became appropriate and would also increase regulatory capacity by increasing government access to and ability to process relevant safety information.

Better information-sharing within government is also critical, especially if a given government generally takes a sectoral rather than a centralized approach to AI governance. Coordinated whole-of-government responses to emergencies are likely to require a degree of coordination between agencies with varying expertise and resources that will be impossible to set up on short notice unless some sort of framework for efficiently and securely distributing sensitive information is already in place and in use. We hope, and even expect, that no dramatic regulatory response will ever be required on short notice. However, the very factors that make governments bad at regulating emerging technologies—the rapid pace and unpredictable course of technological development—also mean that there’s no way to be sure significant state intervention won’t suddenly become vitally important at some point in the future, as it sometimes has in the past. The better informed the government is, the less heavy-handed its response in such emergency situations will have to be.

Flexible Rules and Definitions

The importance of flexible, adaptable rules to the effective governance of emerging technologies is well established. Prematurely implementing rigid rules increases the risk of misspecification, and flexible rules are less likely to be rendered obsolete by technological progress. There are a number of promising ways to make AI regulations more flexible; for instance, regulators could implement if-then commitments (conditional rules that go into effect only upon the occurrence of a specified trigger condition) rather than traditional prescriptive regulations. Alternatively, regulators could take a management-based approach under which AI companies would be required to take risk mitigation measures but given broad discretion over what measures to implement and how.

Another crucial component of regulatory flexibility is the creation of flexible definitions. For example, many recently proposed AI laws apply only to "frontier models" or an equivalent term in order to regulate the most capable (and therefore potentially dangerous) models without imposing compliance costs on smaller startups. However, defining "frontier model" (or "frontier company," if the law in question is entity-based rather than model-based) often proves a more difficult task than you might expect. Early efforts generally relied on a compute threshold, i.e., on a measurement of the amount of computation used to train a model. But statutory definitions that rely solely on a simple compute threshold will become obsolete fairly quickly, as low-compute models become radically more capable and as the cost of compute decreases in accordance with Moore's law. This being the case, it's generally a good idea to preserve optionality by leaving the task of defining "frontier model" to a regulatory agency that can promulgate and regularly update a regulatory definition. This is because regulatory definitions can be updated more rapidly and reliably than definitions baked into statutes.

California's SB 1047 illustrates the importance of this point. If SB 1047 had not been vetoed, it would have placed requirements on "covered models," and defined "covered model" to include only models that cost in excess of a hundred million dollars to train. This cutoff would have made sense in the summer of 2024, when SB 1047 was passed, but it would have been rendered more or less obsolete mere months later in January 2025 by the Chinese startup DeepSeek, which apparently spent less than six million dollars on the

final training run for the base model on which the state-of-the-art reasoning model DeepSeek-R1 was built. Amending the statutory cost threshold in SB 1047 to account for new developments would have required passing an entirely new bill, a costly and time-consuming process that would have taken many months if it could be accomplished at all. A regulatory definition, by contrast, could have been updated much more quickly in response to developing circumstances, and therefore would face a reduced risk of being rendered obsolete by technological progress.

There are tradeoffs involved, of course. Earlier versions of SB 1047 contemplated a more easily updated regulatory definition of "covered model" that did not include the cost threshold; the cost threshold was likely introduced in order to provide assurance to startups and other stakeholders that only very large businesses would be affected. More generally, giving regulatory agencies unsupervised control over questions of immense economic and political significance can be constitutionally problematic, and agencies are generally viewed as being less democratically accountable than legislatures. It's not at all unreasonable to be worried about giving unaccountable government bureaucracies significant authority over the development of an era-defining technology. But the optionality-preserving value of having flexible rules for a technology with such importance to national security is worth some tradeoffs. Democratic and constitutional concerns can be addressed, at least partially, by limiting the scope of the authority delegated to agencies and by ensuring that the relevant agencies are responsive to congressional and White House oversight.

The EU AI Act provides another case study on the importance of flexibility. The Act has been praised for its inclusion of sophisticated updating mechanisms, specific regulatory review requirements, and other adaptability-increasing features. However, while the Act "delegates substantial authority to amend annexes and procedural regimes, … core definitional frameworks remain frozen." For example, certain provisions of the Act affect only "general-purpose AI models with systemic risk," and the task of defining this category is largely left to the discretion of a regulator (the European AI Office). But one aspect of the definition that was not entrusted to the discretion of the AI Office is the focus on models rather than on, for example, the companies that make them. That choice will be difficult to

revisit in the future, regardless of whether the best solution ultimately turns out to be model-, entity-, or use-based regulation.

Assessments and Evaluations

Governments around the world generally seem to realize the importance of evaluating and assessing frontier models. Like information-gathering, government evaluation of models is both instrumentally and intrinsically valuable; in addition to producing potentially valuable information about existing models, it provides an opportunity for agencies to develop expertise in conducting and securely sharing information about assessments and evaluations. In addition to evaluating models directly, governments should also encourage and subsidize the creation of a vibrant third party evaluations ecosystem, in which independent organizations like METR and Apollo Research would work closely with labs to assess the safety of new models.

No one really seems to disagree (publicly, at least) with the notion that we need to invest more in testing and evaluations. OpenAI recently published a set of policy proposals that includes some extremely sensible suggestions about expanding CAISI's role in conducting evaluations and setting standards. Anthropic has made similar suggestions in the past. So industry seems to be on board, as is civil society, and President Trump's AI Action Plan devoted an entire section to the importance of "Build[ing] an AI Evaluations Ecosystem." But legislative proposals aimed at establishing a federal evaluations regime, like the Artificial Intelligence Risk Evaluation Act introduced by Senators Hawley and Blumenthal, have been treated as messaging bills without a serious chance of being enacted. If society is going to be prepared for the arrival of transformative AI, that needs to change. The first step, in the U.S., is simply to codify and fund CAISI with an expanded mandate.

Securing Model Weights and Algorithmic Secrets

Over the last couple of years, there's been increasing recognition of the national security importance of securing frontier AI labs against the theft of valuable secrets by foreign adversaries and other bad actors. Aschenbrenner dedicates an entire essay to this topic in Situational Awareness, concluding that security at frontier AI companies is currently so poor that there's little realistic chance of protecting model weights and algorithmic secrets from foreign adversaries (China, in particular) in the face of a serious and sustained attempt to acquire them.

The U.S. Department of Defense recognizes AI development as an emerging technology area of particular importance to U.S. national security. It follows that lab security is an important national security issue as well. Securing access to frontier AI research is critical to retaining optionality; obviously, a given government's ability to retain the option of exercising control over the way that systems developed within its borders are used depends to a great extent on whether foreign adversaries can simply make off with the model weights and other algorithmic secrets used to develop said systems. In the event that a closed-weight model demonstrates capabilities with a high degree of national security relevance prior to being deployed, a lab with secure model weights can simply choose to delay public deployment until appropriate security measures are implemented. If a foreign power or nonstate actor already has access to the model weights, however, the capabilities in question will already be theirs to do with as they please.

There are a number of measures that governments should take as soon as possible to preserve optionality without imposing a significant regulatory burden on companies. Perhaps most importantly, the U.S. federal government should promulgate comprehensive voluntary standards for physical and cybersecurity throughout the frontier AI development supply chain. Some work has been done in this direction already, but, as the Trump administration's AI Action Plan acknowledges, a significantly more detailed and comprehensive effort is needed. One promising way to encourage compliance with these standards, once they exist, is to make compliance with them a condition of federal grants and contracts, as the Department of Defense currently does with its Cybersecurity Maturity Model Certification program.

Securing information against state-level espionage efforts is one area where certain government institutions have significantly more expertise than just about any private entity. Some of the tech companies developing frontier models have experience securing sensitive information, but none of them has multiple decades of experience administering a program as complex as the U.S. government's classification system for national security information and the related system of security clearances for regulating access. The government also has access to intelligence information about, e.g., the objectives of foreign governments that may need to be shared with frontier AI companies in certain situations. Setting good standards, and creating efficient and secure channels for sharing security information, should therefore be viewed as a national security priority.

Hiring and Talent

The most important factor in building governmental capacity for AI governance is access to top-tier talent. Ideally, governments should be focused on acquiring *more* elite talent than is required to meet the demands of the current regulatory landscape, because the existence of a deep reserve of talent increases optionality. If and when dramatic regulatory action is required—if, for example, it suddenly became clear that a frontier model posed a significant threat to national security, as contemplated in AI 2027—having the necessary personnel ready in advance would enable governments to respond quickly and decisively.

Currently, however, the governments of both the EU and U.S. are struggling to hire highly qualified employees to fill critical AI-related positions. Many factors contribute to this ongoing failure: private sector compensation packages can exceed government salaries by orders of magnitude, government hiring processes in the U.S. are outdated and inefficient, and legislatures and agencies generally lack the sense of urgency that the political moment calls for. This makes it difficult for governments to acquire the kind of deep technical expertise necessary to effectively govern frontier AI systems.

In both Europe and the U.S., most stakeholders will, if pressed, agree that governments need to get better at hiring AI expertise. The Trump Administration's AI Action Plan repeatedly emphasizes the importance of

acquiring and retaining talent, and suggests a number of specific policy initiatives for furthering this goal. The Biden Administration also repeatedly emphasized the importance of federal hiring of AI talent, e.g. in § 10.2 of executive order 14110. The UK's AI Security Institute (AISI) has made hiring AI researchers from frontier labs a key performance indicator in both the Sunak and Starmer governments. But even the UK, which has provided its world-leading AISI with roughly ten times as much funding as its U.S. counterpart receives, isn't spending anywhere near the amount that a rational government should if it accepts the likelihood or even the possibility of transformative AI systems being developed in the next decade. AISI's current job postings offer between £65,000 and £145,000; even the UK government's more aggressive No. 10 technology recruitment program has advertised salaries only up to £200,000—a mere fraction of the compensation packages that the private sector offers to talented AI researchers.

Increasing the funding of CAISI, the AI Office, and relevant sectoral regulatory agencies and enabling them to pay employees more is necessary, but not sufficient. New hiring and contracting authorities and creative approaches to the use of existing authorities are needed to ensure that the EU and U.S. governments have access to talent and expertise. In the U.S., this could mean establishing a "reserve corps" of private-sector experts who could be called in to advise the government in an emergency, or reforming the Intergovernmental Personnel Act to allow it to be used to draw on private-sector AI talent. In the EU, it could mean reducing bureaucratic delay in the hiring process, pushing back against political pressure to ensure proportional representation for all of the EU's member state nationalities, and making fuller use of the AI Act's Scientific Panel of Independent Experts to bring technical expertise into the implementation process without relying exclusively on civil-service hiring.

Avoiding Premature and Overbroad Preemption

Ultimately, regulating frontier AI systems in the United States should primarily be the responsibility of the federal government. Frontier models are used throughout the country in interstate commerce, and regulating them well is a complex task that only the federal government can realistically develop the capacity to undertake. Uniform federal requirements are, in principle, preferable to a patchwork of state regulations imposing inconsistent requirements on AI companies and burdening innovation. At some point, it will almost certainly become necessary for the federal government to preempt some state AI regulations in order to make room for federal regulation. The question on which substantial disagreement exists is not whether preemption should ever occur, but rather *when* and *how* preemption should be implemented.

Radical optionality provides a useful perspective on this problem. Consider the moratorium on state AI regulation that was introduced as part of the June 2025 reconciliation bill before being stripped out in the U.S. Senate. This moratorium went through several versions, but in essence it would have prohibited states from enforcing any law regulating "AI," broadly defined, for ten years.[4] In our opinion, this moratorium was ill-advised for a number of reasons, including, notably, its effect on optionality.

It is difficult to pass federal legislation in the United States. In 2023, for example, Congress passed only 27 bills that became law. Practically speaking, this means that a broad preemption bill would be both (a) unlikely to be reversed, once passed, and (b) unlikely to be followed by any substantial federal AI legislation in the near future. In other words, preempting state AI legislation and replacing it with nothing radically reduces the available regulatory option space. State transparency bills such as SB 53 in California and New York's RAISE Act are a second-best option; ideally, they would be replaced by a single uniform set of federal transparency requirements. But eliminating the second-best options without replacing them with any federal framework, while acknowledging that a federal framework is necessary, would be foolish.

Concerns about the potential for a burdensome patchwork of state regulations governing frontier AI development are reasonable, but they are also almost purely hypothetical. Currently, the development of frontier AI

systems in the United States is almost entirely unregulated. Eliminating states' ability to regulate because some state regulations *might* prove burdensome in the future would be unwise. Instead of locking the country into five or ten years without the possibility of state AI regulation based on the assumption that such regulation will burden innovation, Congress should wait until it becomes clear which kinds of state regulation are unnecessarily burdensome and preempt those laws. Instead of *predicting*, the federal government should *react*, and make any binding, committal decisions under as little uncertainty as possible. It will always be possible to preempt state laws once a federal standard rendering them redundant has been passed, and if the federal government decides that certain kinds of regulation should not exist on either the state or federal level, Congress can pass a deregulatory preemption bill along the lines of the Airline Deregulation Act of 1978.

Of course, the difficulty of passing federal legislation means that the iterative process recommended above will be hard to pull off. But preemption is only rarely accomplished by a standalone bill—more often, preemption measures are packaged along with affirmative federal laws. This means that a one-to-one approach—for example, passing federal transparency requirements and preempting state transparency requirements in the same bill—is realistic, as well as being the historical norm. Additionally, authorizing one or more agencies to regulate AI will automatically give those agencies the authority to regulatorily preempt state laws that intrude on federal regulatory authority. And it makes far more sense to allow state governments to regulate by default, rather than prohibiting all regulation, in the absence of future federal AI legislation. Federalism is an important value in our constitutional system, and the sovereign right of states to pass laws protecting the safety and welfare of their citizens shouldn't be lightly cast aside.

Essentially, radical optionality suggests that we should wait to preempt state AI laws until we know what we're preempting and why. This means that the scope of any preemptive measures should be narrower than the scope of the recent moratorium, and that preemption should take place *after* a federal approach to a given regulatory problem has been decided, not before. If the federal approach is deregulatory, well and good—but that decision should be made *before* a given area of state AI regulation is preempted. Any other approach risks losing out on any benefits that could arise from state

regulation. And, given that states are capable of actually passing AI legislation, while the federal government has thus far proven incapable of doing the same, those benefits are far from negligible.

It's worth noting that this approach—narrow rather than broad, iterative rather than ex ante—is the only approach to preemption of state laws regulating an emerging technology that has ever been taken in the history of the United States. The 2025 moratorium's approach, preempting nearly all state laws regulating a given technology before any federal law regulating the technology had been passed, was unprecedented. In the past, preemption of state regulation of emerging technologies has always occurred after the federal government had decided how the technology should be regulated. This allows the federal government to preempt only those state laws inconsistent with the federal scheme that has been decided on, while leaving laws within traditional areas of state authority unaffected. It also allows the states and the federal government to iteratively work through the complex question of how regulatory responsibilities should be distributed. Trying to answer this question before the contours of the problem are even clear is like trying to put together a jigsaw puzzle while blindfolded.

Objections

If you've made it this far, you probably have a number of bones to pick with our proposal. The remainder of the essay is devoted to briefly responding to a few criticisms that we think get to the crux of the debate.

Objection #1 — Giving the Government a Hammer

It could be argued that increasing the capacity of government agencies to take dramatic regulatory action will increase the likelihood of dramatic regulatory action being taken, perhaps prematurely or unwisely. Giving regulators hammers, in other words, might lead them to hallucinate nails. This is a legitimate concern, and we can't promise that there's no chance of new authorities being abused, or that everyone will agree on when dual-use AI systems have become so advanced that regulating them is a national security imperative. That said, attempting to prevent overregulation by intentionally hamstringing regulators seems short-sighted. Instead, we should guard against the risk of overzealous regulators by ensuring that democratically elected branches of government have meaningful oversight capabilities. In the U.S., this means congressional and White House oversight of agencies, and in the EU it means meaningful parliamentary oversight of the European AI Office and other AI-relevant regulatory bodies.

Consider the specific affirmative interventions recommended in the previous section: information-gathering authorities, whistleblower protections, risk assessments and evaluations, personnel recruitment measures, hiring and contracting reforms, and measures to increase lab security. These aren't weighty substantive authorities that lend themselves to abuse, they're common-sense ways of increasing regulatory capacity so that action can be taken when and if the appropriate authorities decide that action is needed. Of course, the potential for abuse isn't zero; mandatory reporting requirements, for example, could become burdensome and harm innovation if overly onerous, and overly broad whistleblower protections might make it easier for employees to make off with valuable trade secrets, harming employers without providing a substantial public benefit. But preventing this from happening is a matter of designing rules well and ensuring that elected officials exercise appropriate oversight. This won't necessarily be easy, but if the problem is taken seriously we think it should be possible.

Objection #2 — Democratic Legitimacy

It could also be argued that preserving optionality might come at an unacceptable cost to democratic legitimacy. Flexibility and democratic legitimacy are often in tension. For example, authorizing federal agencies in the U.S. to waive notice and comment requirements for rulemaking allows agencies to respond quickly to technological developments, but also eliminates an avenue for public input into the rulemaking process. Likewise, delegating rulemaking authority to agencies preserves optionality because agencies can update rules more rapidly and frequently than Congress can update laws, but the Supreme Court has recently issued a number of decisions limiting Congress's ability to delegate this kind of authority to agencies because of concerns about the democratic and constitutional implications of agency rulemaking.

At the end of the day, a balance has to be struck between democratic responsiveness and flexibility. This doesn't mean that democratic legitimacy is less important in the AI governance context than elsewhere. In fact, the opposite is true. Because powerful AI systems threaten to "reshape the delicate balance between state capacity and individual liberty that sustains free societies," potentially pushing democracies towards despotism, it's vitally important for the processes by which AI is governed to be (and to be seen to be) legitimate and responsive to public concerns. But failing to adequately prepare government institutions increases the risk of extremely undemocratic outcomes, like the "Manhattan Project" scenario that Aschenbrenner predicts in "Situational Awareness." Increasing optionality, by building up regulatory capacity and ensuring that the government and the public are well informed, is a way of increasing the odds that there will be other viable options and that society will have the time and resources necessary to choose between them in a democratically legitimate way.

Objection #3 — Concentration of Power and Government Abuses

A synthesis of the previous two objections might go something like: "you want to give governments an arbitrary amount of capacity to regulate this transformatively useful general-purpose technology, but what if they use all that capacity to do bad things, like entrench themselves, remove democratic guardrails, and illegally punish political enemies?"

This is a point that deserves to be taken quite seriously. Much has been written about so-called "concentration of power" risks arising from governments asserting control over the development of highly capable future AI systems, and we're basically convinced that the threat model discussed in that literature should impose serious constraints on the nature and extent of the authorities we give to our governments.

This, in fact, is why we didn't include "radically expand the scope of emergency authorities like the Defense Production Act" as one of our policy suggestions. Emergency authorities were for many years (and still are, to some extent) a central focus of the AI governance literature, because the classic catastrophic risk scenarios ("loss of control" over misaligned superintelligence; terrorists using advanced AI systems to design apocalyptically virulent bio-weapons) require quick and decisive executive action above all else. But emergency authorities have been a double-edged sword since antiquity, and as the capabilities of the most advanced AI systems have continued to rapidly improve, the other edge of the blade has increasingly come into focus. Concentrating power in the hands of a few is a paradigmatic loss of optionality—it forecloses future choices for everyone else.

The ongoing dispute between the Pentagon and Anthropic illustrates the problem nicely. On one hand, the government appears to be abusing authorities intended to be used against foreign adversaries and in wartime emergencies in order to punish a domestic company for perfectly legal positions taken in a contract negotiation. On the other hand, the actions taken or threatened (shutting a given AI company out of defense industry supply chains; forcing an AI company to modify a given model to prevent unacceptable behavior) are exactly the kinds of actions that one might want the executive branch to be able to take, swiftly and without friction, in a genuine emergency.

There is no solution to this age-old problem, but there are better and worse approaches to addressing it. One particularly promising approach would require governments to use only law-following AI systems. This might involve creating benchmarks and standards to measure how effective the guardrails built into advanced AI systems are at preventing the system from agreeing to carry out an illegal order, and then passing legislation prohibiting government officials from using systems that fail to perform satisfactorily.

The more fundamental point here is that not all optionality-increasing policies contribute, on the margin, to concentration of power risks. There are many policies that would increase governments' capacity to keep their citizens safe without symmetrically increasing the same governments' capacity to threaten citizens' safety. It's possible to imagine situations in which a government agency being extremely good at (for example) conducting safety evaluations of frontier models might contribute to bad outcomes, but in our judgment those situations seem quite unlikely.

Objection #4 — Private Governance Is All You Need

Dean Ball has argued that AI governance should depend on private governance mechanisms "just as much, if not more than" government laws and regulations. This seems probably correct, but (as Ball acknowledges) it does not mean that government institutions have no role to play.

The two most noteworthy proposals for private governance regimes have been Dean Ball's Framework for the Private Governance of Frontier Artificial Intelligence and Gillian Hadfield and Jack Clark's Regulatory Markets: The Future of AI Governance. Both of these papers make a number of excellent points about the importance of private governance mechanisms to AI regulation and suggest interesting solutions to some of the most obvious obstacles to private governance. But a fundamental feature of both proposals is a highly competent and well-resourced government office charged with overseeing the private regulators and ensuring (either through ex ante licensing decisions or ex post license revocations) that only competent and aligned entities are allowed to regulate. In other words, both proposals acknowledge that private governance alone is insufficient, and that government will inevitably play a key role in even the most free-market AI governance regime.

The primary legislative proposals attempting to put something like a private governance regime into effect have focused on setting up "independent verification organizations," or IVOs. A detailed discussion of the pros and cons of specific state IVO bills is beyond the scope of this paper; suffice it to say that we think the basic idea is sound and that we would wholeheartedly support a well-scoped, optionality-enhancing IVO proposal. Our contention is simply that no system for private governance, however well designed, will

eliminate the need for capacity-building efforts within government. At the very least, the critical task of licensing and/or supervising private governance entities will need to be addressed by a government office.

The advantages of private governance mechanisms in the AI governance context are numerous: they are more flexible, they can take advantage of the resources and personnel of the sophisticated actors developing frontier AI models, and it will likely be easier to incorporate AI tools into private governance institutions than into government agencies. Because of all this, it makes sense for the time being to limit the active role of public institutions mostly to coordinating and facilitating communication between private stakeholders. Once the development of truly transformative dual-use AI systems is imminent or underway, however (an eventuality that seems to grow nearer with each passing month), some degree of actual regulation may become necessary, because old-fashioned top-down government regulation has a few fundamental advantages that will be difficult, if not impossible, to replace.

Since transformative AI would profoundly affect the lives of everyone in the world, it simply isn't acceptable that important decisions about how its benefits and risks should be weighed should be left solely to the companies creating it. This isn't a matter of trusting or not trusting AI developers, it's simply a recognition of the reality that private for-profit companies should, do, and in many cases are legally obligated to prioritize the interests of their shareholders above the interests of the general public. Attempts to work around this fact with clever corporate governance structures have so far failed spectacularly. Luckily, it's very likely that the interests of the relevant shareholders will be aligned with the interests of the public to a great extent, but the extent to which they are aligned depends on the existence of capable public institutions.

For example, one of the main factors disincentivizing companies from creating AI systems that might cause harm to large numbers of people is the fact that doing so would likely result in significant tort liability for the company. It isn't an indictment of the personal character of any individual employee of an AI lab to say that, in the absence of tort liability, it is unlikely that their employer would adopt the socially optimal level of precaution against such risks. The beauty of a competitive market economy is that it forces companies to operate efficiently, and to respond to whatever

incentives exist, or to go out of business. This being the case, we should ensure that our existing tort law systems are designed in such a way as to align the interests of frontier AI companies with the interests of the public to the greatest extent possible. It's not at all clear that currently existing systems of tort law does this, so it's important for legislatures to start implementing sensible reforms as soon as possible.

Another issue with placing complete responsibility for self-regulation on AI companies is that the enforcement mechanisms available to them are necessarily more limited than those available to governments. Say, for example, that a consortium of AI companies establishes a set of best practices for safely developing highly capable models. Given the information and resources available to the consortium, it's very plausible that these standards would be better than anything the government could have come up with on its own. However, if at some point in the future it becomes clear that it is an urgent national security imperative that *all* AI companies in a given country should abide by a particular standard, the consortium will not have the legal authority to force any defectors to do so. Simply trusting that no company will act recklessly and fail to comply with a purely voluntary standard when there are many billions of dollars on the line is a recipe for disaster. While it would be unwise to hamper innovation by regulating prematurely on the assumption that such a situation will arise, it would be equally unwise to fail to prepare for the possibility altogether.

In an ideal world, perhaps there would be some way around the need to involve the government in AI governance at all—but, as leading AI companies mostly seem to acknowledge, there isn't. No amount of ingenuity in designing novel private governance structures and mechanisms is likely to create an adequate substitute for governments' monopoly on violence, or to confer democratic legitimacy on private corporations that do not and should not primarily serve the public interest. It's true that government institutions are often cumbersome, bureaucratic, and slow to act, but this just means that laying the groundwork to minimize delay and maximize competence is all the more important.

Conclusion

Recall the four basic assumptions from the beginning of this essay. We assume that there is a real possibility that transformative AI will be developed within the next decade; that profound uncertainty exists regarding what capabilities and characteristics the technology will have and what the best way to govern it will be; that as a dual-use technology with significant national security implications, it will eventually require some degree of government oversight; and, crucially, that building the capacity required to provide that oversight will take time.

If you generally accept these premises, you should support radical optionality. Instead of relying on uncertain predictions about exactly how transformative AI will be developed and what it will be capable of, governance efforts should focus on ensuring that, when and if important decisions need to be made, governments have the institutional capacity to make them well. This can be accomplished by taking steps that will be helpful regardless of when and how transformative AI is developed, without burdening innovation to any significant extent.

The concrete measures we've discussed—information-gathering authorities, whistleblower protections, secure information-sharing channels, flexible rules and definitions, robust assessments and evaluations, lab security standards, and hiring and talent reforms—are first steps toward that goal. More ambitious proposals, including large-scale public investment in compute, may also be justified, but our focus here has been deliberately pragmatic. In other words, the policies we have proposed have the advantage of being realistically achievable now, both in the U.S. and in Europe. The cost of implementing these policies is modest, relative to the potential benefits. The cost of failing to act, by contrast, is potentially catastrophic.

Endnotes

1 Regulation isn't the only factor that has contributed to the relative stagnation of Europe's tech industry, and may not even be the primary cause, although it very likely plays a role. See Ian Hogarth, *Can Europe build its first trillion-dollar start-up?* https://www.ft.com/content/e3e23aea-eb4d-42dd-a9ea-9ae267b8f507.

2 "Dual-use" technologies have peaceful civilian applications, but can also be used for military, terroristic, or otherwise harmful purposes. For example, advanced semiconductors can be used in a variety of harmless civilian products (such as video game consoles) but also in missile guidance systems. The term is most often used in the context of export controls. The U.S. government has in the past described advanced general-purpose AI models with certain potentially dangerous capabilities as "dual-use foundation models."

3 Of course, this benefit hasn't always resulted from market forces alone. From time to time, government intervention to address harmful externalities has played an important role, as when the depletion of earth's ozone layer was successfully reversed as a result of the Montreal Protocol, an international agreement that limited the production of harmful chlorofluorocarbons.

4 The final version of the moratorium would have lasted for only five years and exempted some state laws from preemption, including "generally applicable laws." It's not entirely clear which laws would ultimately have been deemed "generally applicable" and which would not. This final version also applied only to states that accepted certain federal broadband infrastructure funds, and allowed states to opt out by refusing the broadband funding.