POLICY BRIEF

A Framework for U.S. AI Governance:

Creating a Safe and Thriving AI Sector

Dan Huttenlocher
Asu Ozdaglar
David Goldston

In consultation with the ad hoc committee on AI regulation

Version of November 28, 2023
(Initial draft for circulation, September 26, 2023)

MIT Schwarzman College of Computing
MIT Washington Office
aipolicy.mit.edu

I. OBJECTIVES

This policy brief is motivated by two objectives:

- Maintaining U.S. AI leadership – which is vital to economic advancement and national security – while recognizing that AI, if not properly overseen, could have substantial detrimental effects on society (including compromising economic and national security interests).
- Achieving broadly *beneficial* deployment of AI across a wide variety of domains. Beneficial AI requires prioritizing: security (against dangers such as deep fakes); individual privacy and autonomy (preventing abuses such as excessive surveillance and manipulation); safety (minimizing risks created by the deployment of AI, particularly in already regulated areas such as health, law and finance); shared prosperity (deploying AI in ways that create broadly accessible opportunities and gains from AI); and democratic and civic values (deploying AI in ways that are in keeping with societal norms).

II. FUNDAMENTAL PRINCIPLES

With those objectives in mind, we offer the following discussion and guidance regarding how AI governance might proceed. Making AI systems subject to a mix of regulation and liability law is important for ensuring that AI is developed and deployed in ways that facilitate its beneficial uses now and over the long run. Users (and the public at large) should have a clear idea of what they are getting (and not getting) from an AI system, and should be able to feel confident about that understanding through mechanisms such as contractual relationships, disclosures, and audits.

To accomplish that:

- *The norms, regulations, and institutions to shape AI need to develop in tandem with it.* Because technology tends to move quickly, existing oversight or regulatory approaches should be extended, where possible, to align AI with current norms and laws (in contrast with de novo approaches to AI regulation such as those being pursued in the EU).
- In other words, *the first step in AI governance should be to ensure that current regulations apply to AI, to the greatest extent possible. If human activity without the use of AI is regulated, then the use of AI should similarly be regulated.* The development, sale and use of AI systems should, whenever possible, be governed by the same standards and procedures as humans acting without AI in the AI's domains of applicability (e.g., in healthcare, law, hiring and finance). Among other things, this will inherently ensure that many higher risk applications

of AI are covered, since those are areas for which laws and regulations have already been developed. (One issue will be how to deal with laws that rest on "intent." AI does not currently have intention, although those developing and using it may.) This approach of applying existing laws will also prevent AI systems from being used to circumvent existing regulations. This type of approach is already being used in some domains; for example autonomous vehicles are held to the same standards as those operated by humans. (General purpose AI like GPT-4 or chatbots built on it like ChatGPT are dealt with further below.)

- This approach of extending current legal frameworks to activities involving AI should also apply to activities undertaken by governments, including policing, setting bail, hiring, etc.
- In general, laws and regulations regarding AI should be promulgated and enforced by the same entity or entities that govern human actions without AI in the same domain. This may require such entities to develop some AI expertise, as discussed further below.
- As AI laws and rules develop further, differences between AI and human capabilities can be taken into account. AI will have capabilities humans lack – for example, in finding patterns or taking into account counter-factual propositions. This may warrant holding the use of AI for some purposes to stricter standards than are applied to humans acting without AI.
- For the approach outlined in this section to succeed, *providers of AI systems should be required to identify the intended purpose(s) of an AI system before it is deployed.* Regulatory agencies might issue guidance or rules to define how intended uses should be described and/or the courts might develop case law that defines what disclosures of intended purpose are adequate to satisfy particular objectives. The latter path, while leaving more ambiguity and risk, might allow providers of AI and their customers to work out what kinds of disclosures are needed, and could leave more flexibility as AI evolves. But that will work only if steps are taken to ensure that those purchasing and using AI have the capability to make effective demands on AI providers and can be sure that any contractual terms they negotiate will be enforceable. Moreover, the more open-ended approach may be more appropriate for some uses of AI than others. Different states might also experiment with different approaches in application domains that are regulated at the state level.
- It is important (but difficult) to define what AI is, but often necessary in order to identify which systems would be subject to regulatory and liability regimes. The most effective approach may be defining AI systems based on what the technology does, such as "any technology for making decisions or

recommendations, or for generating content (including text, images, video or audio)." This may create fewer problems than basing a definition on the characteristics of the technology, such as "human-like," or on technical aspects such as "large language model" or "foundation model" – terms that are hard to define, or will likely change over time or become obsolete. Furthermore, approaches based on definitions of what the technology does are more likely to align with the approach of extending existing laws and rules to activities that include AI.

- *Auditing regimes should be developed* as part and parcel of the approach described above. To be effective, auditing needs to be based on principles that specify such aspects as the objectives of the auditing (i.e., what an audit is designed to learn about an AI system, for example, whether its results are biased in some manner, whether it generates misinformation, and/or whether it is open to use in unintended ways), and what information is to be used to achieve those objectives (i.e., what kinds of data will be used in an audit).

  There are several possible ways (not necessarily mutually exclusive) an effective auditing "ecosystem" could develop. For some uses, the government (federal, state or local) might require that an audit be conducted (by a provider, regulatory entity, and/or user) before and/or after a system is deployed, particularly in higher risk uses such as application domains that are already regulated (e.g., healthcare). Or, a system might develop more organically, with users demanding audits or conducting them (before or after deployment), and with courts assessing liability more severely when an AI system does not perform as claimed, or violates a regulation, and no appropriate audit has been conducted. Audits might be carried out by third parties, by users, or by the government. Any audit would have to protect the intellectual property of the AI system provider and the user.

  Audits can be readily manipulated to not reveal problems that exist or conversely to suggest something is problematic when it is not, so public standards for auditing will need to be developed. Such standards might be established by a nonprofit entity, for example one analogous to the Public Company Accounting Oversight Board (PCAOB), or by (or with) a federal entity like the National Institute of Standards and Technology (NIST).

- It is important to understand the limitations and requirements of different kinds of audits. For example, prospective audits allow a system to be tested before it is put into use, which can be useful in high-risk settings. But the distribution of data in such *a priori* testing may be quite different from that encountered in actual use (e.g., an audit of an AI system that will be used to identify good candidates for

employment may use a different distribution of candidates than those who will appear in the actual applicant pool). Retrospective audits, by contrast, may require access to information that is confidential to the organization being audited, so it will be important to establish clear guidelines and accountability for keeping data confidential.

- "Explainability" – automated explanations of how an AI system reached a conclusion or generated a result that are both accurate and understandable to humans – is not currently possible and may never be. But *AI systems can, and should be more interpretable*, i.e., provide a sense of what factors influenced a recommendation or what data was drawn on for a response. Increased interpretability should be encouraged through means that are not highly prescriptive about the technology because methods will continue to evolve quickly. The government might encourage greater interpretability through a regulatory system that places more requirements on AI systems that are less susceptible to interpretation, or courts may develop a more severe liability regime for such systems.
- Training data play a central role in many AI systems. Many data sources, particularly from the public internet, contain inaccuracies, biases, inadvertent disclosures of private information, and various defects that AI systems should be designed to guard against. Clear statements of intended outcomes for AI systems, along with mechanisms such as testing, monitoring, and auditing (both for specific application domains and general use) are important, in part, to limit problems that can arise from problems in the training data.

III. AI STACKS FOR SPECIFIC APPLICATIONS

- It is important to recognize that AI systems will often be built "on top of" each other; for example, a general-purpose system like GPT-4 might be the foundation of another system that is used in hiring. We think of this as an "AI stack" comprising any AI system that can serve as a component of other systems, including but not limited to foundation models. (This is analogous to general software systems, which can comprise many components connected via program interfaces (APIs) and are commonly referred to as a "software stack.") *In general, the ultimate provider and user of a service provided by an AI stack would be responsible for how it operates*, but in doing so would need to depend on components of the overall stack performing correctly. So, for example, a stack deployed for hiring would primarily be the responsibility of the provider and user of the full stack (and the provider would be subject to the regimes described above). But when a component system of a stack does not perform as promised, it may be reasonable for the provider of that component to share responsibility.

To facilitate that, those building on systems that depend on general-purpose AI should seek information on how the general-purpose system would handle the particular intended function and ask for enforceable guarantees that the general-purpose system will perform as stated. (As noted above, this presupposes a legal environment where contractual guarantees can be demanded and enforced.) It should be noted that AI components in a stack may interact in unexpected ways, making it important to consider all components in a stack together –for example by auditing such systems (both prior to and perhaps after deployment).

- A goal (and test) of a regulatory and liability system is that it should, in effect, clarify what constitutes a *"fork in the toaster" situation* – that is, when a user (rather than an AI provider) is responsible for a problem because the AI system was used in a way that was clearly not responsible or intended. Such situations can be identified only if providers spell out what the proper uses are, if there are best-practice guardrails against other uses, and if the legal responsibilities of providers are clearly delineated and widely understood. In addition, the uses and limits of AI need to be broadly understood by users. By analogy, one can't be held personally responsible for putting a fork in a toaster, if neither the nature of toasters nor the dangers of electricity are widely known. What a reasonable user, or a reasonable developer of an application, is expected to know is likely to change over time. *The AI system provider should in most cases be held responsible* for a problem unless it is able to show that a user should have known that a use was irresponsible and could not have been foreseen or prevented by the provider.

IV. GENERAL PURPOSE AI

- Providers of general-purpose AI systems like GPT-4 (and broad applications built on them like ChatGPT) cannot be expected to disclose all the intended uses of their systems. Yet these systems carry additional risks because of their broad uses, broad availability and the potential for applications such as chatbots to interact in ways that feel human to users. Despite the broad domains of potential use of general-purpose systems, *the government could require general AI systems to disclose whether certain uses are intended* (such as dispensing medical advice) as well as to have guardrails against unintended uses (which could be identified in regulation). Regardless of disclosures of intent, general AI systems – like all others – would have to be in compliance with laws and regulations governing activities that would be covered for humans acting without AI. (See above.) In addition, statutes and/or case law could impose more severe liability for problems that arise from uses that the provider could reasonably have foreseen but did not prevent with guardrails or provide sufficient warnings or

instructions about to the user. Any warnings or instructions should be provided in a manner that a user is likely to see and heed. Pages of small print that a user is unlikely to read are not adequate.

- *Providers of general-purpose AI might be required to monitor uses of their AI systems* (as pharmaceutical companies are often required to monitor effects of their products after they are in the marketplace). This will be especially important if general-purpose AI systems are developing new capabilities as they are used (rather than just changing when an updated system, like GPT-4, is released.) The government could specify the kinds of problems that providers are responsible for uncovering, reporting and addressing. The government could also require that new general-purpose AI systems be piloted before being made widely available.
- Note, what is considered "general-purpose AI" will likely change over time as the technology develops, and regulation will need to be flexible enough to evolve with the technology.
- There are areas in which regulation regarding AI should go beyond what is typically imposed on human actors, for example where AI has capabilities that people do not. One such area is using AI to create fictitious but realistic-seeming images, video or audio involving the likeness of actual places and people (alive or dead), sometimes called "deep fakes," (especially when used without proper permission). While people have long been able to manipulate images, AI makes it as easy to create a realistic fake as it is to take a real photo and also makes it easier to target recipients of disinformation. Another such area is privacy and surveillance. While pre-AI technologies were capable of violating individual privacy and conducting intrusive surveillance, these capabilities are significantly amplified by advances in AI. In addition, *the government should require images that were created by AI to be clearly marked as such* – both through labels that can generally be seen by humans and through machine-detectable means, such as watermarking.

V. AN AI REGULATORY AGENCY

- For oversight regarding AI that lies beyond the scope of currently regulated application domains, and that cannot be addressed through audit mechanisms and a system similar to that used for financial audits, *the federal government may need to establish a new agency that would regulate such aspects of AI*. The scope of any such regulatory agency should be as narrow as possible, given the broad applicability of AI, and the challenges of creating a single agency with broad scope. The agency could hire highly qualified technical staff who could

also provide advice to existing regulatory agencies that are handling AI matters (pursuant to the bullets above). (Such a task might alternatively be assigned to an existing agency, but any existing agency selected should already have a regulatory mission and the prestige to attract the needed personnel, and it would have to be free of political and other controversies from existing missions that could complicate its oversight of AI.) A self-regulatory organization (like the Financial Industry Regulatory Authority, FINRA, in the financial world) might undertake much of the detailed work under federal oversight by developing standards and overseeing their implementation.

- Even with a new AI agency, regulation of AI systems for specific uses (such as aiding in medical diagnoses) would continue to reside in the existing agency governing that domain (such as the Food and Drug Administration), as described above.

VI. OTHER ASPECTS OF GOVERNANCE

- *A goal of any regulatory or liability approach should be to encourage more research in the private sector, universities, and research organizations on how to make AI systems beneficial* (as defined in the Objectives section of this report). Just as regulation of the automotive sector has led car companies and universities to conduct research on how to make vehicles safer and less polluting, AI regulation should lead companies to do more research on how to make AI systems more beneficial (including, for example, by making them more interpretable). In addition, the government needs to increase its funding of public research on these issues. Of particular importance is public research on large-scale AI systems, such as large language models, upon which other AI systems are built.

- The development of beneficial AI systems *requires a clear framework for intellectual property (IP) rights*. It is important that human creative endeavors remain incentivized in a world with AI systems. While court decisions are affirming that authors and inventors must be human, so no IP rights accrue to AI, questions remain about how IP laws apply in the context to AI systems (such as copyright, which gives creators of a work rights as to how the work is distributed and used as part of other works). In particular, while it has always been possible to create works that infringe on copyright, AI-generated content seems poised to greatly increase this. (Note, whether something constitutes infringement is ultimately determined through case law and litigation.) Uncertainty over how creators can guard against and readily identify potential infringement, as well as over which people or organizations would be responsible, are complicating the use of generative AI. The situation is changing rapidly. For instance, Microsoft

announced that for paying customers of its generative AI products, the company will defend and pay any adverse judgments or settlements for copyright infringement generated by its products, as long as the customer used the guardrails and content filters built into the products. There are also questions regarding the use of digital personal data that are much broader than AI (such as in social media) that should be approached in a manner that does not constrain AI more than it does other technologies.

- AI systems are currently trained under the assumption that copyright does not prevent such use of materials, but this is being challenged by content owners. Such questions frequently arise with new technologies. A relevant example is the Google Books case, filed in 2005 claiming copyright infringement by Google in scanning a large number of books to make them searchable online. In 2015, the courts upheld a district court decision that this did not constitute infringement because material was not being reproduced. Rather, the resulting search tool provided information as to where copyrighted material could be obtained with permission. While the training of AI systems, similar to the Google books case, does not directly produce any content, an AI system differs in that its use may generate infringing content. It is currently unclear whether AI-generated infringing content will be easier or harder to identify than human-generated infringing content. On the one hand, there is the likelihood of more infringing content being generated, but on the other, a copyright holder could in principle prompt an AI system in order to demonstrate that it produces infringing content, without needing to undertake a search for specific instances of possible infringement. It is also possible that AI systems might be developed that provide better reference to what original sources are relevant to a given output, helping evaluate the possibility of infringement.

- Open source can both help prevent problematic uses of AI – by allowing more individuals to uncover flaws – and increase such uses – by making it easier to build AI systems that can cause harm (including by bypassing guardrails) and making their development and distribution harder to control. *Developers of open source models should follow the same practices as commercial providers, in clearly establishing intended uses and following best practices for guardrails against other uses.* Providers of services that use open source AI as components need to assume responsibility for how that component behaves. While these are analogous issues to the use of open source software, the greater challenges entailed in interpreting and understanding AI systems compared to standard software may increase risks.

In summary, it is unclear whether and how current regulatory and legal frameworks apply when AI is involved, and whether they are up to the task. This leaves providers, users and the general public in a *caveat emptor* situation. There is little to

deter the release and use of risky systems and little incentive to proactively uncover, disclose, or remediate issues. Those who build on general models have insufficient information about the component systems of their stacks and little recourse if problems result. Additional clarity and oversight regarding AI are needed to facilitate the development and deployment of beneficial AI and to more fully and smoothly realize AI's potential benefits for all Americans.

This paper was prepared by Dan Huttenlocher, Dean of the MIT Schwarzman College of Computing; Asu Ozdaglar, Deputy Dean of the MIT Schwarzman College of Computing and Head of the Department of Electrical Engineering and Computer Science; and David Goldston, Director of the MIT Washington Office, with guidance from an ad hoc committee on AI regulation they assembled. The committee members are:

Daron Acemoglu, Institute Professor, Department of Economics

Jacob Andreas, Associate Professor, Department of Electrical Engineering and Computer Science

David Autor, Ford Professor of Economics

Adam Berinsky, Mitsui Professor of Political Science

Cynthia Breazeal, Dean for Digital Learning and Professor of Media Arts and Sciences

Dylan Hadfield-Menell, Tennenbaum Career Development Assistant Professor of Artificial Intelligence and Decision-Making

Simon Johnson, Kurtz Professor of Entrepreneurship, MIT Sloan School of Management

Yoon Kim, NBX Career Development Assistant Professor, Department of Electrical Engineering and Computer Science

Sendhil Mullainathan, Roman Family University Professor of Computation and Behavioral Science, University of Chicago Booth School of Business

Manish Raghavan, Assistant Professor of Information Technology, MIT Sloan School of Management

David Rand, Schell Professor of Management and of Brain and Cognitive Sciences, MIT Sloan School of Management

Antonio Torralba, Delta Electronics Professor of Electrical Engineering and Computer Science

Luis Videgaray, Senior Lecturer, MIT Sloan School of Management