Insight

Primer: AI Governance Frameworks

JOSHUA LEVINE | JUNE 21, 2023

Executive Summary

- The proliferation and popularity of artificial intelligence (AI) systems such as ChatGPT-4, DALLE-2, and Stable Diffusion have raised questions about what AI is, how it will impact people's lives, and whether it is safe.
- Congress and the Biden Administration, as well as their international counterparts, have begun to propose and enact various frameworks to address the potential benefits and harms of AI-powered products and services; these frameworks generally fall into one of two models – a flexible, pro-innovation approach or a precautionary approach.
- This primer discusses what is driving the interest in AI, what regulatory frameworks are being proposed, and how different proposals overlap and diverge.

Introduction

Since ChatGPT's release in late November 2022, artificial intelligence (AI) and its potential have sparked conversations around the world over how this technology should be regulated. Business leaders and policymakers are increasingly interested in the capabilities of new AI-powered products and services and the tradeoffs this technology presents. While AI is not necessarily a "new" technology, a combination of advances in hardware, software, and data allow engineers to create systems that perform a host of functions – from text editing, image and video generation, and computer code generation, to network management – more efficiently and accurately than ever before.

These advances have led to a wide range of responses about, and proposals for, the future of AI. For some, AI tools and services may be the engine of the next great productivity boom and usher in a new age of technological innovation. For others, AI presents an existential threat that could further entrench existing inequalities and create new threats, with some

going as far as to say AI could lead to human extinction.

Some frameworks to regulate AI, such as the Organization for Economic Cooperation and Development (OECD), the United Kingdom's (UK) pro-innovation approach to AI regulation, and the National Institute for Standards and Technology (NIST) Risk Management Framework (RMF) embrace a more flexible, pro-innovation approach that emphasizes the rapidly evolving capabilities of AI systems, the use of existing legal tools to address potential harms, and a better understanding of the tradeoffs created by more stringent regulation. Contrasting these approaches are the frameworks proposed by the European Union (EU) and the Chinese Communist Party's (CCP) Cybersecurity Administration of China (CAC), which emphasize compliance with existing state laws related to privacy, discrimination, and acceptable content, require licenses to develop and make AI systems publicly available, and embrace a statutory and rigid approach for compliance and enforcement. Somewhere in between is the White House AI Bill of Rights, which places heavy emphasis on AI's potentially harmful impact on protected classes and democracy and outlines how developers, deployers, and regulators should work to mitigate risks before a system is adopted, while recognizing the important role the private sector will play in driving innovation and access to these technological capabilities. Congress held three hearings focused on AI in May, another two hearings in June, with a field hearing in Silicon Valley pending. President Biden recently traveled to San Francisco to meet various AI researchers and advocates. Members of both chambers have introduced several pieces of legislation that would respectively create a federal AI task force, provide AI training for federal employees, deny AI firms Section 230 immunity with regard to generative AI, and secure the software supply chain for the Department of Defense.

This primer discusses recent developments in AI, where these technologies are already having an impact and potential areas of future disruption, and how domestic and international policymakers and organizations are thinking of regulating AI.

What Is AI?

AI refers to the use of computers and machine learning to mimic the problem-solving and decision-making capabilities of the human mind. AI's components are software, specifically algorithms, hardware, microchips such as semiconductors and graphic processing units, and data. While forms of AI have been used for decades to help process and sort information and make decisions, advancements in technology are creating new opportunities for firms and individuals to leverage AI in a wide range of uses. AI is currently used in mobile apps such as Google Maps and Spotify to give directions and music recommendations, or to play games such as IBM's "Deep Blue," which defeated chess grandmaster Garry Kasparov in 1997. AI systems can better detect and understand diseases such as cancer, allow

autonomous vehicles and drones to navigate roads and the national airspace, and improve the way students learn. Further, the release of Large Language Models (LLM), such as ChatGPT and Claude, and generative AI models such as Stable Diffusion, provide users with access to tools that help them learn, work, create and communicate with technology like never before.

On the other hand, AI presents potential risks to individuals and society, such as entrenching existing inequalities and exacerbating harms, which have led to public concerns related to job displacement, predatory data collection, individual privacy, cybersecurity, and monopolization. Further, the "black box" nature of certain AI models, a term used to describe when developers or deployers cannot explain how a system produces an output, contributes to concerns that AI could drive unequal treatment for different populations in health care, employment, and financial opportunities. On the more extreme end, there have been calls to halt development in cutting edge AI models due to the threat of a mass-extinction event. Such concerns have led to government officials in the United States and around the world to consider regulatory guardrails for the use of this technology.

Flexibility vs. the Precautionary Principle: Different Frameworks for AI Regulation

Any regulatory model should seek to maximize benefits and limit harms related to AI adoption and innovation. How institutions craft regulations to solve this tradeoff is where divergence begins. Generally, this divergence has produced two paths.

Some institutions propose a flexible, non-statutory approach that relies on *ex post* regulation to address potential future harm. This approach tends to embrace AI innovation, and attempts limit potential barriers to new entry as more firms leverage the capabilities of AI in different contexts. This approach to AI regulation, largely influenced by the OECD, is embodied in the UK's pro-innovation policy and the NIST RMF. The Biden Administration's blueprint also incorporates many of these ideas, though to a lesser extent. The flexible approach to regulation allows for regulatory principles to be updated and iterative, building on research and experience agencies, industry, and civil society generally gain over time. A more flexible approach also allows for growth and development of new AI models but runs the risk of allowing harms to accrue before regulators have a chance to identify and resolve them.

Some institutions take a much more restrictive path that focuses on pre-empting harm through regulation, regardless of the effect on innovation and development of new AI models. Primarily embodied in the EU AI Act and the CCP's CAC guidelines for deep synthesis technologies and generative AI, these types of regulation tend to require innovators to receive permission to deploy new technology and seek to pre-emptively

regulate the entire AI development life cycle in a top-down fashion. By acting preemptively, regulators can take more control in identifying and eliminating potential harms before they can develop, but at the cost of innovation and the deployment of new services.

One example to illustrate the real-world application of these principles is the use of an LLM to help with customer service for a cable company and with screening patients in a doctor's office. Both scenarios focus on interactions between people and are likely augmented by existing technologies such as an automated answering system or an online portal. A call center could use an LLM to provide more accurate responses to customer queries, improving response time and accuracy for the caller, and improving the call-center employees' job performance. Potential harm would likely be minimal and limited to a problem canceling or switching an account or improper charges.

Using such AI in the medical context, a patient could report their symptoms in a portal that is used to prompt a model analyzing the symptoms and the patient's medical history to provide potential diagnoses. This could eliminate time spent by nurses and doctors on intake and allow providers to spend more time with patients, but the model could lead to an incorrect diagnosis, which could compound existing problems or create new ones.

With a more flexible framework, firms would grapple with existing sector-specific regulations related to sensitive financial or medical information and equal treatment for protected classes, but generally allow firms to experiment and integrate new technology to better serve their customers. Harms to customers could be addressed through existing consumer protection law or laws related to medical malpractice. Under the EU or Chinese models, both AI programs would need to be screened and licensed by various regulators before deployment and any potential violations could result in revocation of the license, removing the system from use, and additional penalties for developers and deployers. Further, in the health care context, additional documentation and impact assessments would likely be required to receive and retain approval for use. Such an approach could minimize harm, but it would be much more difficult for consumers to realize the benefits of AI.

Regulatory Frameworks and How They Diverge

While some general trends regarding flexibility and control can be identified, the frameworks do vary even within these broader categories. Understanding these differences can help policymakers craft effective AI policy.

OECD Recommendation on Artificial Intelligence

The OECD recommendations published in May 2019 promote standards that are flexible and

can evolve over time to keep pace with the rapidly changing AI landscape. These recommendations include five principles for the responsible stewardship of AI, including inclusive growth, human-centered values, transparency and explainability, safety and security, and accountability. The OECD emphasizes the importance of international co-operation and trust, in addition to national rules to promote cross-border collaboration, innovation, and a policy environment that enables rather than restricts innovation. These principles and ideas are present and cited in several of the proposals including the UK report, NIST framework, and to a lesser extent the White House blueprint, which all emphasize sector-specific expertise to guide flexible regulation. Further, these three approaches recognize the important role the private sector will play in both innovating and addressing harms and seek to involve them in crafting solutions rather than treat them as a threat to be controlled. The driving theme of the OECD's principles is flexibility to ensure the recommendation can endure as technology and applications continue to evolve.

U.K.: A Pro-Innovation Approach to AI Regulation

The UK's secretary of state for science, innovation and technology presented a report to Parliament in March 2023 on plans to support innovation while providing a framework to address and mitigate risks. The report emphasizes using existing regulations to ensure developers and deployers are free to innovate as well as build public trust and competency to address potential harms in the future. Much like the OECD framework, the UK approach seeks to balance growth and innovation with safety, transparency, and accountability for development and use of AI systems. The UK report caveats any harm reduction principles with the goal of promoting innovation and ensuring small and medium firms are not uniquely disadvantaged by regulation. Further, the report recommends "central support functions" the government could conduct to promote domestic innovation and mitigate harms, such as "horizon scanning and gap analysis" to evaluate potential threats, creating testbeds and sandboxes led by various agencies, and promoting interoperability among British and international standards.

NIST AI Risk Management Framework (RMF)

The NIST framework, published in January 2023, targets adequate risk management with a focus on flexibility similar to the OECD recommendations, but pays greater attention to varied levels of risk posed by different types and uses of AI systems. Keeping with the OECD framework, NIST's RMF highlights that different types of AI systems present different risk considerations and regulations should be proportionate to potential harms. NIST guidance calls for measures related to transparency, accountability, and reliability to mitigate potential harms and promote trust between developers, deployers, and impacted individuals. Both the NIST RMF and UK report, however, advocate more aggressively for

governance structures that promote innovation than the Biden Administration's blueprint.

EU Artificial Intelligence Act

The EU Parliament is considering a draft compromise that would represent the world's first statutory framework for AI. The EU approach is heavily focused on the potential harms AI systems present, including general purpose AI, but specifically "high-risk" use cases that could impact health, safety, fundamental rights, the environment, as well as political campaigns and social media recommender functions for firms covered under the EU's Digital Services Act. The bill would require any products that could meet these criteria to go through assessments and receive approval from individual member states' and the EU government's regulators before being made available to the public with the potential of being revoked at any time. These provisions are similar to those outlined by the CAC related to testing and transparency, risk assessments, and more stringent rules for sensitive information such as biometrics. They diverge, however, with requirements that outputs conform to a specific political ideology. In the EU, there is a focus on ensuring AI does not undermine democratic values and elections, whereas CAC regulations ensure AI promotes CCP values such as socialism and protects the "national image" and "social public interest."

Provisions on the Administration of Deep Synthesis Internet Information Services & Administrative Measures for Generative Artificial Intelligence

The CCP's CAC published a framework for "Deep Synthesis of Internet Information Services" in December 2022, and in April 2023, began a proceeding to regulate generative AI. Both frameworks outline specific requirements for the types of systems developers can build, the types of data they can use, restrictions on the dissemination of certain information, and an expectation that all systems will uphold and advance socialist values. The EU AI Act and the CAC framework do consider different use cases carrying different levels of risk, but unlike other approaches, require extensive pre-deployment vetting as well as specific compliance mechanisms throughout the AI life cycle. The former CAC framework ascribes regulatory authorities to CCP agencies, as well as lays out rules governing the use of "deep synthesis technology," which includes technology for generating or editing text, images, simulations, and text-style conversations, specifically detailing what types of content and actions are prohibited under Chinese law. This includes disclosing information on who uses the technology, how algorithms are written, which types of data can and cannot be used, and what outputs of systems will be allowed. The latter framework deals specifically with generative AI and follows many of the same themes laid out in the former provision.

The Biden Administration's Blueprint for an AI Bill of Rights

The Biden Administration's blueprint for an AI bill of rights focuses heavily on the potential risks automated systems and AI could pose to Americans and provides principles to guide development and deployment of AI intended to maximize benefits while minimizing harms. The document seeks to pre-emptively address harms from AI by setting principles for development, use, and post-deployment monitoring. The blueprint recommends proactive equity and disparity assessments prior to and following deployment of AI systems, while also elevating existing regulatory tools to address harms. In comparison, the NIST RMF focuses on mitigating harms, but is more flexible in how it seeks to minimize such harms and highlights innovation as a key attribute of the framework. This includes soliciting input from impacted communities in the training process; ensuring training data is high-quality, representative of the population, and does not reinforce existing inequities; and ongoing monitoring and mitigation after a system is put into use. The blueprint lays out important documents and frameworks that influenced its creation as well as principles related to data privacy and engaging relevant stakeholders when considering further regulation.

Conclusion

While automated systems and forms of AI are not new, the latest developments with LLMs and generative AI mark a new phase of development and innovation. This new phase has the potential to create both significant benefits and harm for businesses, governments, and individuals. Governments and organizations across the globe are attempting to craft frameworks that maximize AI's benefits and minimize its risks according to their own interests and goals. Their frameworks illustrate some of the similarities and differences of national values related to innovation, regulation, and the role of the state in guiding technological development and innovation.