Carnegie Mellon University
Responsible AI

The Block Center
FOR TECHNOLOGY AND SOCIETY

JANUARY 2023

Toward AI Accountability:
Policy Ideas for Moving Beyond a Self-Regulatory Approach

Summary

The U.S. Government must strengthen its commitment to the responsible development and use of Artificial Intelligence (AI) by implementing more holistic and proactive regulatory policies and developing new industry incentives and enforcement mechanisms. Responsible development and use of AI has transformative potential to enhance human capabilities, bolster economic growth, and increase quality of life for all. However, existing efforts to build responsible AI systems have been largely reactive and ad-hoc in nature. Enhanced USG efforts should complement and advance the mission of the National AI Initiative and Federal agencies to lead the world in the development and use of trustworthy AI in the public and private sectors. Such efforts must articulate that economic growth can be achieved through AI systems that provide real benefits to end-users while attending to issues of equity and harm reduction.

- The fast pace of AI systems' advancement and accelerating deployment across industry demand a more active Federal role to ensure its responsible development and use. Enforcement agencies must strengthen emerging norms and apply clearly-articulated policies to ensure compliance and accountability for these systems.
- While industry will continue to play a key role in developing norms and institutionalizing best practices regarding the development and implementation of accountable AI systems, effective legislation should acknowledge — but surpass — a self-regulatory approach, which tends to address harms after they are realized.
- A dependable system of checks and balances regarding the use of AI technologies will confer competitive advantage to nations applying such measures, engendering trust and reliance in responsibly-developed AI systems.

Risk of Inaction

A lack of clear Federal guidance regarding the accountable development and use of AI systems will exacerbate and amplify the harmful biases and risks to safety, privacy, equity, agency, and security that these systems exhibit in the real world. Lack of clear guidance will also perpetuate incentives to overstate performance beyond system capabilities. These cumulative risks carry long-term consequences if left unaddressed.

- Existing policies and legal frameworks fail to enforce an adequate degree of accountability because current legislation is *domain-specific, focused on outcomes without adequate consideration of processes, and reactive in nature.* As with the regulation of traditional software products and services, current governance mechanisms do not offer clear, precise process guidelines that businesses can adhere to. This gap has provided some organizations room to exploit regulatory ambiguity, increasing the risk of harmful AI.

Challenges for Effective Regulation

AI systems pose intrinsic accountability challenges. Machine learning (ML) — the branch of AI that has seen rapid technological advancements and applications in the last decade — can extract statistical patterns from data to produce predictions that now surpass human capabilities in an increasing number of tasks. However, the highly complex process by which an ML system produces predictions often requires intensive empirical verification to evaluate claims of benefit and to detect the sources of error in systems.

- The highly complex nature of the trained models makes it difficult and time-consuming to fully understand why the system makes certain errors or finds particular patterns. These models cannot be easily translated or explained with semantically meaningful language.
- As in traditional software development, error attribution for ML-based systems also encounters the "many hands" problem, where true accountability for a suboptimal output could reside simultaneously at many stages of development or deployment, including the procurement of AI systems, data curation, statistical modeling, and embedding of the algorithm in a larger socio-technical system. Effective accountability frameworks must fully consider all stages of the design and deployment of the entire system.
- Foreseeing the long-term harms of these systems can be challenging, because such harms are often diffuse in nature and dynamically evolving. Examples of such harms include accessibility challenges for new AI systems, and possible environmental consequences associated with the use of some AI technologies. However, the use cases of ML in high-stakes domains are ever-expanding, and likely will require novel systems of both individual and corporate accountability.

Defining Artificial Intelligence Broadly

This memo adopts the definition of artificial intelligence (AI) offered by the National Institute for Standards and Technology's draft AI Risk Management Framework, which refers to engineered or machine-based systems designed to operate with varying levels of autonomy that can, for a given set of human-defined objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments.

Policy Recommendations

Require meaningful involvement from a wider set of stakeholders in the design, evaluation, and deployment of AI systems to account for new risks, to ensure that benefits are realized for key stakeholders, and to reconceive an accountability framework.

- The Federal government must support research efforts toward designing and testing methods that involve the widest possible set of stakeholders and appropriate domain expertise in the design, evaluation, and deployment of AI systems. Such participatory approaches can close the communication gap between developers and impacted individuals and communities and ensure that the resulting AI systems respond to real-world needs and risks.
- Participatory methods should facilitate meaningful stakeholder participation in the process of choosing the overall objective of the AI system, the development of the model, an evaluation of the AI's behavior, and evaluations to ensure the intended outcomes for which the AI was developed are realized.

Ensure that Federal regulations scrutinize both the development process and the product, ideally through separate evaluation mechanisms. An improved evaluative regime must also enable a shift from a component-level to a system-level view of accountability.

- Among other criteria, implementing agencies should require developers and practitioners to use understandable objectives and models, and clearly scope the intended usage of their systems.
- Existing regulations in this area — for example, the European Union's General Data Protection Regulation (GDPR) — require impact assessments to be carried out prior to the uptake of AI technologies. U.S. regulators must follow suit, moving towards common Federal standards across agencies, developing concrete impact assessment standards and templates, and mandating such assessments prior to the deployment of AI.

Policy Recommendations (continued)

Encourage the development of industry best practices, develop new incentives for industry compliance, and clearly articulate and enforce regulations. A greater USG focus on norms development and compliance could incentivize well-established corporations to strengthen internal accountability processes. However, effective Federal guidelines must strike an appropriate balance between economic growth and intellectual property rights on the one hand, and minimizing the subsequent harms of AI systems on the other. This can be achieved through effective auditing and placing the burden of producing reliable evidence on industry.

- **Develop clear standards for articulating intended use cases** and metrics for reliability and utility so that users can have clear expectations of performance under well-defined conditions.
- **Define limitations on the use of AI technologies** by creators and users. Recommendations include limiting liability to harms resulting from intended use as defined by the AI developer, limiting the use of very large datasets and defining appropriate cases for their development and management, and defining how datasets and systems that have significant impacts on populations are managed.
- **Define test and evaluation frameworks, measurement and metrics, and continuous monitoring standards** based on the assessed risk of the application space or use case, and developer choices with regard to data, benchmarks, metrics, models, and overall implementation. Compliance with regulatory audits above a defined risk threshold should be mandatory but highly protected to preserve industry IP and sensitive data.
- **Develop ratings and indices capturing compliance** with Federal guidelines that, for example, could be publicized on a regular basis to ensure transparency and raise public awareness of compliance within industry.

Policy Recommendations (continued)

Properly resource the relevant U.S. Government agencies to meet this evolving challenge. A lack of appropriate technological expertise will hinder the effective enforcement of any updated regulatory regime. Supporting broader USG efforts to leverage, hire, and retain the necessary expertise is critical to any effort to define and enforce AI accountability in industry.

- Congress must prioritize funding for expanding use-case inspired and policy-relevant research in AI fairness, accountability, transparency, and ethics. Given the challenge of matching private-sector salaries in attempts to recruit AI and data scientists into public service, Federal grants in this area could be stipulated on researchers' commitment to working with USG agencies in some defined but limited capacity to support the development and enforcement of AI accountability measures in industry.
- Federal resources must be allocated to develop the tools and infrastructure necessary to help scale compliance enforcement, and train auditors capable of understanding and utilizing these state-of-the-art tools and protocols to effectively align AI technologies with consumer needs and broader societal values. The USG must also articulate good compliance metrics and ensure effective program management when awarding USG contracts to determine industry compliance.
- Implementing agencies such as the Federal Trade Commission (FTC), and standards-setting agencies such as the National Institute of Standards and Technology (NIST), must also fully leverage creative hiring programs such as the U.S. Digital Corps and the U.S. Digital Service to develop the necessary levels of AI and data science expertise.

This policy brief was drafted and published by affiliated faculty and staff of the Responsible AI Initiative and the Block Center for Technology and Society at Carnegie Mellon University.

If you would like to schedule a briefing with Carnegie Mellon faculty experts to discuss AI accountability, toolkits to address bias in AI systems, or a deeper dive on policy ideas conveyed in this memo, please contact responsibleai@andrew.cmu.edu.

Co-authors listed in alphabetical order: Brandy Aven, Justin Deyo, Motahhare Eslami, Jodi Forlizzi, Sarah Fox, Rayid Ghani, Hoda Heidari, Kenneth Holstein, Christian Kaestner, Ramayya Krishnan, Norman Sadeh, Carol J. Smith, Molly Steenson, John Zimmerman