PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

via www.regulations.gov

National Telecommunications and Information Administration
AI Accountability Policy Request for Comment
Docket Number: NTIA-2023-0005

June 12, 2023

We are researchers associated with the Center for Information Technology Policy (CITP) at Princeton University and write to highlight a few areas where the NTIA can help guide the development of trustworthy Artificial Intelligence ("AI") systems.[1]

The NTIA's request for comment ("RFC") sets forth a rich set of questions about how to develop an AI accountability system. We offer three core principles that the NTIA should take into account in developing recommendations for future regulation. *First*, the accountability ecosystem should have multiple, overlapping mechanisms for ensuring that AI systems are serving the public interest. *Second*, because AI systems involve complex socio-technical interactions between data, models, and people operating in different institutional contexts, assessments cannot look at one element in isolation to form a judgment about the whole system. *Third*, while many current assessment tools focus on important questions about whether AI systems are biased or unfair, it is equally important to assess whether the AI

[1] In keeping with Princeton's tradition of service, CITP's Technology Policy Clinic provides nonpartisan research, analysis, and commentary to policy makers, industry participants, journalists, and the public. These comments are a product of that Clinic and reflect the independent views of the undersigned scholars in computer science, public policy, and law.

systems are fit for purpose. In particular, many systems are used by organizations to make consequential decisions about individuals that are based on unreliable science and make dubious claims of fairness, accuracy, or efficiency.

This area is rapidly developing and so our comments are necessarily preliminary in nature. Nevertheless, we detail four potential avenues for the NTIA to promote mechanisms that improve accountability: (a) enabling the development of a standards-setting body; (b) focusing on whether AI systems are fit for purpose and based on rigorous science; (c) requiring that AI systems can be examined from multiple vantage points; (d) developing oversight mechanisms for public sector use of AI systems. We welcome the opportunity to continue participating in the discussion around how AI systems should be regulated to serve the public interest.

A. Accountability Objectives (RFC Q1-8)

1. Assessment Standards

One important function the NTIA can play in its report is to clarify the purpose and scope of different assessment mechanisms. Researchers and practitioners in the AI accountability field use terms such as "audits" or "impact assessments" without much consensus about what the different mechanisms involve.[2]

Traditionally, audits in the financial sector have involved a standard setting body developing guidelines for how to assess the financial disclosures of a business, and then an independent, certified professional firm examining that business to see

[2] *See Examining the Black Box Tools for Assessing Algorithmic Systems*, Ada Lovelace Institute, https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

how it fares under those standards.[3] The goal of a financial audit is to give investors assurance that they have high quality information about the business, which in turn aids the public trust in the capital markets. Audits cover both governance controls and metrics for reporting financial information. There are efforts underway in the sustainability field to develop parallel standards for environmental, social, and governance risks.[4]

Drawing on these experiences, the NTIA can support the development of a consensus based standard setting organization to identify the purposes of an AI audit, who are the relevant stakeholders, what information is required to assess if an AI system is reliable, and what metrics are helpful to stakeholders.[5] As the experience of the sustainability and financial standard setting bodies have shown, such standards will need to be grounded in sector specific considerations.

There are also a variety of accountability mechanisms that use risk assessments (pre-deployment) and impact assessments (post-deployment) to measure whether an AI system operates in an unfair or discriminatory fashion. Such assessments can be conducted externally by treating the system as a "black box," or can be run using internal data and privileged access to the model. As discussed in more detail below, these different mechanisms provide different insights about the AI system across its lifecycle.

---

[3] *See, e.g.*, Paul Munter, *The Importance of High Quality Independent Audits and Effective Audit Committee Oversight to High Quality Financial Reporting to Investors*, https://www.sec.gov/news/statement/munter-audit-2021-10-26

[4] *See The Sustainability Reporting Ecosystem*, https://sasb.org/about/sasb-and-other-esg-frameworks/

[5] The American Association for Public Opinion Research has a self-regulatory initiative designed to bring more openness around their research methods for surveys that highlights the value of transparency for value-based judgments. https://aapor.org/standards-and-ethics/transparency-initiative/.

2. Fit for Purpose

A specific gap in accountability studies is assessment tools for whether an AI system is actually fit for purpose. This is particularly important in applications of AI for predictive optimization: automated tools that make decisions about individuals based on predictions about their future outcomes.[6] Predictive optimization is a distinct type of automated decision making that has proliferated widely. It is sold as accurate, fair, and efficient. There are dozens of applications of predictive optimization already in use, including in consequential domains such as criminal justice and child welfare. In our recent research, we found that applications of predictive optimization suffer from severe flaws that challenge the legitimacy of these applications.[7] To hold developers of such applications to account, we provide a rubric of specific questions that developers should adequately address before they can deploy an application of predictive optimization.[8] The risk is that absent rigorous testing and validation, the AI system can be used in consequential settings with surprisingly low accuracy and impact vulnerable populations.[9]

3. Reproducibility & Validation

A related area that is understudied presently is that many of the purported advances in machine learning are difficult to reproduce and externally validate. As

---

[6] Angelina Wang, Sayash Kapoor, Solon Barocas, & Arvind Narayanan, *Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy,* Presented at ACM FAccT 2023. https://predictive-optimization.cs.princeton.edu/

[7] *Id.*

[8] *See* https://predictive-optimization.cs.princeton.edu/rubric.pdf

[9] *See* Matthew Salganik et al., *Measuring the predictability of life outcomes with a scientific mass collaboration*, Proceedings of the National Academy of Sciences (2020) https://www.pnas.org/doi/10.1073/pnas.1915006117.

a result, AI systems can be deployed in high-stakes scenarios where they have not been properly vetted. Several researchers have documented failures to reproduce prominent scientific findings in machine learning.[10] In recent research from CITP, we have found that a leading cause of reproducibility failures is data leakage: when data used for training an AI model is also used for evaluating it.[11] Data leakage is widespread across fields: it affects hundreds of papers across dozens of disciplines. In the absence of systematic interventions, leakage will continue to lead to irreproducible research, and could ultimately reduce trust in scientific research that uses AI. One possible intervention is to use model information sheets.[12] But there are other avenues to consider as well. The practical concern is that absent reproducible research, it might be impossible to investigate previous behaviors of the model and to know if it is producing reliable results.[13]

4. Generative AI

In the case of generative AI systems that use large language models ("LLMs"), a recent paper lead by a CITP-affiliated researcher and other co-authors proposes a three-layered approach, whereby governance, model and application

---

[10] Michael Roberts et al. *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans*, Nature Machine Intelligence (2021), https://www.nature.com/articles/s42256-021-00307-0; Giles Vandewiele et al. *Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling*, Artificial Intelligence in Medicine (2021), https://www.sciencedirect.com/science/article/abs/pii/S0933365720312525.

[11] Sayash Kapoor and Arvind Narayanan. *Leakage and the Reproducibility Crisis in ML-based Science*, Arxiv (2022), https://arxiv.org/pdf/2207.07048.pdf.

[12] *See* https://reproducible.cs.princeton.edu/model-info-sheet-template.docx.

[13] For example, a generative AI system might demonstrate surprisingly high performance in certain tests because it has memorized training data that was leaked to it during development. *See* Arvind Narayanan and Sayash Kapoor, *GPT-4 and professional benchmarks: the wrong answer to the wrong question*, https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks.

audits inform and complement each other.[14] During governance audits, technology providers' accountability structures and quality management systems are evaluated for robustness, completeness, and adequacy. During model audits, LLMs' capabilities and limitations are assessed along several dimensions, including performance, robustness, information security, and truthfulness. Finally, during application audits, products and services built on top of LLMs are first assessed for legal compliance and subsequently evaluated based on their impact on users, groups, and the natural environment.

B. Improving Transparency (RFC Q9)

There are a number of approaches that are used currently to assess AI systems. The mechanisms differ based on how the information about how the system operates is collected from the AI system. In some instances, data is crowdsourced, or collected through other means that treats the AI system as a black box. In such approaches, ensuring representative samples or carefully separating the effects of the AI system from other confounding variables or sources is often a challenge.[15]

Other mechanisms depend on varying degrees of privileged access to the data that is released by the system operator. In the overwhelming majority of cases, access to the training data, model (weights), code, and optimization objectives

[14] Mökander, J., Schuett, J., Kirk, H.R., Floridi, L. *Auditing large language models: a three-layered approach*. AI Ethics (2023). https://doi.org/10.1007/s43681-023-00289-2.
[15] Basileal Imana, Aleksandra Korolova & John Heidemann, *Auditing for Discrimination in Algorithms Delivering Job Ads*, Proceedings of the Web Conference, 2021, https://dl.acm.org/doi/fullHtml/10.1145/3442381.3450077/.

underlying the AI system are not made available for external scrutiny. Moreover, in many instances, even access to the outputs of the AI model is limited.

The NTIA should encourage a variety of assessment mechanisms to proliferate because they each present a different picture of how an AI system operates. Some issues can be caught early in the pre-deployment phase, but other issues may only emerge through seeing how the system operates in practice. And outsiders may have a vantage point that demonstrates issues with information that system operators may provide to privileged parties. Like the financial sector, there should be a variety of players with different types of access who have the incentive to detect and expose problems with an AI system.

The NTIA can promote standardized transparency mechanisms to allow external actors to evaluate AI systems. For example, there could be an "inspectability API" requirement that gives researchers access to query the system.[16] While there are potential concerns that privileged access would allow researchers to reverse-engineer such systems or learn personal information, those concerns can be assuaged through appropriate certification of the researchers. But it is equally important that users are empowered with the right to export their data

[16] For example, a recent paper used such an API for studying the societal desiderata of the AI used in relevance estimators of social media platforms. *See* Basileal Imana, Aleksandra Korolova, & John Heidemann, *Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest*, Proceedings of the 26th ACM Conference On Computer-Supported Cooperative Work And Social Computing (2023), https://dl.acm.org/doi/abs/10.1145/3579610; *see also* Matthew Salganik and Robin Lee, *To Apply Machine Learning Responsibly, We Use It In Moderation* (2020), https://open.nytimes.com/to-apply-machine-learning-responsibly-we-use-it-in-moderation-d001f49e0644(using an API to evaluate a machine-learning software used to moderate comments).

and have third parties (with appropriate consent) help them understand how decisions about them are being made in a standardized and reliable manner.

C. Public Sector (RFC Q19 & Q30)

While much of the current public discussion about AI systems have focused on private sector use cases, there are also many consequential public sector applications that rely on AI systems.[17] The public sector should have a higher bar for deploying such systems and they presumptively require democratic oversight. This is equally true for systems built at the direction of government agencies and those that are procured by the government. In fact, the government might use its procurement power to mandate external assessments of AI systems the government acquires. The NTIA can advocate for mechanisms that will allow for democratic oversight and external audits of these systems.[18] In turn, those mechanisms could be adapted for private sector applications.

* * *

To recap, AI systems are complex socio-technical systems that require scrutiny from multiple vantage points to ensure that they are serving the public interest. We commend the NTIA for embarking on this public consultation and look forward to future opportunities to engage in the public deliberation process to help harness the power of AI systems for good.

---

[17] *See* Karen Levy, Klya Chasalow, and Sarah Riley, *Algorithms and Decision-Making in the Public Sector*, Annual Review of Law and Social Science, Vol. 17, pp. 319-334, (2021), https://www.annualreviews.org/doi/abs/10.1146/annurev-lawsocsci-041221-023808.

[18] *See Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, GAO Report (June 2021), https://www.gao.gov/assets/gao-21-519sp.pdf.

Respectfully submitted,

Archana Ahlawat
Emerging Scholar

Justin Curl
Research Assistant

Sayash Kapoor
Graduate Student, Computer Science

Aleksandra Korolova
Assistant Professor of Computer Science & Public Policy

Mihir Kshirsagar
Clinic Lead

Surya Mattu
Digital Witness Lab Lead

Jakob Mökander
Fellow

Arvind Narayanan
Professor of Computer Science

Matthew J. Salganik
Professor of Sociology

Contact:
Website: https://citp.princeton.edu
Phone: 609-258-5306
Email: mihir@princeton.edu