THE
FUTURE
SOCIETY

The Future Society's Response to NTIA's Request for Comment on AI Accountability Measures and Policies

Prepared by The Future Society (TFS)

June 12, 2023

Stephanie Weiner,
Acting Chief Counsel
National Telecommunications and
Information Administration (NTIA)
U.S. Department of Commerce
1401 Constitution Avenue NW,
Washington, DC 20230

Subject: The Future Society's Response to NTIA's Request for Comment on AI Accountability Measures and Policies (NTIA-2023-0005-0001)

Dear Ms. Weiner,

On behalf of The Future Society[1], we are pleased to submit comments in response to the National Telecommunications and Information Administration (NTIA) Request for Comment on Artificial Intelligence ("AI") systems accountability measures and policies.

The Future Society (TFS) is a U.S. non-profit 501(c)(3) organization, with the mission of aligning AI through better governance. We develop, advocate for, and facilitate the implementation of AI governance mechanisms, ranging from laws and regulations to

[1] https://thefuturesociety.org/

voluntary frameworks such as global principles, norms, standards, and corporate policies. Through our activities, we hope to ensure that AI systems are safe and adhere to fundamental human values.

We are deeply concerned by the rate of development and deployment of general-purpose AI systems (GPAIS), which are already exhibiting safety and security vulnerabilities,[2] and posing threats to human safety and fundamental values. We hence appreciate your attention to this topic of importance and urgency.

Based on the analysis of questions presented in the sections below, TFS urges NTIA to consider and support the following recommendations in its forthcoming activities and policies:

I. Strengthen accountability throughout the entire lifecycle of an AI system, with particular scrutiny applied in the design and development stages of general-purpose AI systems (GPAIS).

II. Foster a trustworthy AI assurance ecosystem by promoting third-party assessments and audits, and contestability tools for impacted persons.

III. Apply a horizontal, cross-sectoral approach to federal policies and regulations for general-purpose AI systems (GPAIS).

The sections below, which provide responses to relevant questions from the NTIA's Request for Comment, substantiate our recommendations.

[2] Burgess, 2023. "The Security Hole at the Heart of ChatGPT and Bing." Wired UK; Cox, 2023. "Facebook's Powerful Large Language Model Leaks Online." Venture Beat; Keary, 2023. "How prompt injection can hijack autonomous AI agents like Auto-GPT." Venture Beat.

THE
FUTURE
SOCIETY

I. Strengthen accountability throughout the entire lifecycle of an AI system, with particular scrutiny applied in the design and development stages of general-purpose AI systems (GPAIS).

Question 16: The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that "[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all." [82] How should AI accountability mechanisms consider the AI lifecycle?

We believe that effective AI governance requires a combination of both binding and voluntary mechanisms, including accountability measures and policies. The ultimate goal of such measures and policies should be to affect the behaviors of actors across the AI value chain[3] in a way that ensures that AI systems are legal, effective, ethical, safe, and trustworthy. To achieve this end-state, we believe such *measures and policies should attend to both the technical characteristics of an AI system as well as the broader socio-technical system in which the AI system is embedded.* Focusing narrowly on technical aspects alone would fail to capture the full range of ethical, safety-related, and legal implications of these systems. In comparison, a broader socio-technical lens incorporates organizational and external factors that influence an AI systems' output and impact.

Additionally, we believe such *accountability measures and policies must span the entire AI system's lifecycle.* This is particularly important for general-purpose AI systems (GPAIS),[4] capable of performing tasks across a broad range of domains, and particularly relevant to policy due to their opaque computational processes, market versatility, and rapid uptake by both individuals and businesses (that adapt GPAIS for a wide array of consumer-facing applications). ChatGPT, for instance, was

---

[3] Küspert, Moës, and Dunlop, 2023. "The value chain of general-purpose AI." Ada Lovelace Institute.

[4] Gutierrez et al., 2022. "A Proposal for a Definition of General Purpose Artificial Intelligence Systems."

recently found to be the fastest-growing consumer application in history, amassing 100 million monthly active users only two months after its release.[5]

The AI system lifecycle is complex, often involving a number of actors and spanning from problem definition, data collection and training to deployment and monitoring. It has been described using various terms and phrases, but, in sum, it can be broken down into three overarching stages: i) design, ii) development, and iii) deployment.[6]

Although most risks may not materialize until the deployment stage of an AI system's lifecycle, there are unique opportunities to identify and mitigate many of these risks from earlier stages; hence, we urge NTIA to promote accountability measures and policies starting from the design and development stages.

We believe this holistic approach will encourage AI developers to address risks proactively rather than as an afterthought.

During the design stage, developers determine the context and goals which will underpin the development and deployment of an AI system, and then start gathering and preparing the data to train their system. During these stages, issues of bias in datasets, leading to discriminatory outcomes, have long been documented.[7] In recent years, these issues have expanded as developers compete to create increasingly more capable GPAIS by training datasets that have been assembled on larger scrapes of the internet, including different modalities such as images, videos, and audio files. We are concerned that GPAIS developers are not giving due attention to removing problematic content (such as malign stereotypes, racist and ethnic slurs, and explicit content, among other problematic types) that could result in such discriminatory outcomes.[8] Beyond biased and discriminatory outcomes, low-quality data collection and processing could also have security and

[5] Hu, 2023. "ChatGPT sets record for fastest-growing user base - analyst note." Reuters.
[6] CoE-GSA, n.a. "Understanding and managing the AI lifecycle" in: AI Guide for Government" U.S. General Services Administration.
[7] Buolamwini and Gebru, 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018; Paullada et al., 2020. "Data and its (dis)contents: A survey of dataset development and use in machine learning research." arXiv.; Weidinger et al., 2021. "Ethical and social risks of harm from Language Models." arXiv.
[8] Birhane, Prabhu, and Kahembwe, 2021. "Multimodal datasets: misogyny, pornography, and malignant stereotypes." arXiv.

safety consequences. For example, recent studies have demonstrated that GPAIS—having been trained with datasets that include information about dual-use biotechnology (e.g. synthetic DNA)—could be prompted to assist non-experts in causing a pandemic.[9]

To prevent societal harms that can arise due to poor decisions made during the design stage, *developers should be responsible for ensuring that the data with which they train their AI systems has undergone robust quality control, involving processes to identify and mitigate bias, and to remove potentially harmful concepts.* Accountability measures and policies should aim to prevent the use of data containing harmful biases, and a subset of scientific information useful for engineering acts of terrorism, such as biological and chemical weapons and cyber attacks. If the system has the potential to impact critical areas such as cybersecurity, elections, defense, biosecurity, or nuclear domains, developers should also be subjected to distinctively stringent accountability regimes beginning from the system's design stage.

During the development stage, developers train, evaluate, refine, and safeguard an AI system. This stage is particularly hazardous because it involves the production and storage of an AI system lacking the security features that are implemented before general release, and evaluations with AI systems with unknown, and, at times novel, characteristics.[10] *Developers should be responsible for mitigating and protecting their AI systems from security threats—including adversarial attacks, hacking threats, data theft, and data leaks*. This is of particular concern because developers currently lack shared standards for cybersecurity and physical security (i.e., protecting against access to their models or facilities) as well as standards for process security (e.g., how AI labs should control for insider threats,[11] audits,[12] structured access,[13] etc.).

---

[9] Soice et al., 2023. "Can large language models democratize access to dual-use biotechnology?" arXiv.
[10] Kaplan et al., 2020. "Scaling Laws for Neural Language Models." arXiv.; Ganguli et al., 2021. "Predictability and Surprise in Large Generative Models." arXiv.
[11] Shevlane et al., 2023. "Model evaluation for extreme risks." arXiv.
[12] Raji et al., 2020. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
[13] Shevlane, 2022. "Structured access: an emerging paradigm for safe AI deployment." arXiv.

To mitigate this broad range of risks, *accountability mechanisms should foremost ensure actors throughout the lifecycle are adhering to existing legal and regulatory frameworks and adopting measures that could substantially improve* safety and security during the development stage. In fact, a recent survey with GPAIS developers indicated extremely high levels of agreement towards such measures, including pre-deployment risk assessments, dangerous capabilities evaluations, third-party model audits, safety restrictions on model usage, and red teaming.[14]

Developers and deployers should also be encouraged to engage with a diverse range of future consumers or stakeholders of the technology throughout their AI systems' lifecycles. This collaborative process can contribute to developers' and deployers' understanding of how their AI systems may affect various individuals and communities, as well as identify potential scenarios of misuse. By involving diverse perspectives, developers and deployers can uncover additional biases, ethical and safety concerns, and preempt societal implications that may otherwise go unnoticed.

Accountability measures within the development stage, such as assessments and audits, should be conducted in a regular fashion, at least once a year. Due to their large user base, GPAIS should be subject to more frequent assessments, triggered, at a minimum, by significant updates or modifications to the AI system, significant changes to the data upon which it relies, or the occurrence of notable adverse events or complaints.

To operationalize the notion of continuous assessments, federal policies and regulations should require monitoring and oversight mechanisms that continuously collect and analyze relevant data in order to detect and respond to emergent issues. The results of these assessments should be communicated transparently through mechanisms such as accessible reports, model cards,[15] or public disclosures.

[14] Schuett et al., 2023, "Towards best practices in AGI safety and governance: A survey of expert opinion." arXiv.
[15] Mitchell et al., 2018. "Model Cards for Model Reporting." FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29-31, 2019, Atlanta, GA, USA.

THE
FUTURE
SOCIETY

II. Foster a trustworthy AI assurance ecosystem by promoting third-party assessments and audits, and contestability tools for impacted persons.

Question 5: Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?

Although there are several types of accountability mechanisms, we believe third-party assessments and audits, and contestability tools are two vital components of a functioning AI assurance ecosystem. They promote trust by assuring that AI systems have been scrutinized by a qualified, independent actor, and help inform people of the requirements that AI systems are expected to meet.

Third-party assessments and audits should be conducted throughout a GPAIS' lifecycle, and feature: an interdisciplinary team with subject-matter expertise; an outcome-oriented approach; standardized metrics; and clear documentation requirements. They should be complemented, when appropriate, by binding regulatory requirements.

Assessments can help ensure the AI system's transparency, safety, and fairness by evaluating them throughout the design, development, and deployment stages. Additionally, they can identify and mitigate security threats, biases, goal misspecification, errors, or other issues that could lead to untrustworthy outcomes[16]. They can enable companies to effectively trace back risks and harms to specific nodes of the AI lifecycle and the actors responsible for them. Hence, they can promote a clearer route to rigorous, trustworthy AI development and agile innovation processes.

Assessments and audits can be conducted by third-parties or internally. Although internal AI assessments play an important role in mitigating risks,[17] we encourage

[16] Brundage et al., 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv.
[17] Schuett, 2023. "AGI labs need an internal audit function."

NTIA to not overlook third-party assessments and audits. These refer to the evaluation of an AI system by an independent organization, which has not been involved in the design and development of the object being tested and must not be intended as the eventual user of the system,[18] in an effort to promote a more reliable, unbiased, and trustworthy AI assurance ecosystem.

To be effective, third-party assessment and auditing regimes should (i) avoid transferring liability from the AI developer and operator to the external auditor, (ii) perform a granular assessment across the AI lifecycle, including the design and development stages; and (iii) analyze business-to-business (B2B) practices in off-the-shelf systems purchases, services agreements, and the subsequent use and deployment to consumers.

We stress, however, that *third-party assessment and audits must not be perceived as silver bullets.* They have several limitations, including risks of focusing on processes but neglecting outcomes and being ineffectual if not conducted by domain experts. Furthermore, external audits, in particular, may be subject to liability-washing (companies seeking to conduct external audits with the ulterior motivation of evading liability). Due to these limitations, these mechanisms must be complemented by robust regulatory regimes, including requirements for internal auditing.[19]

In developing assessment and auditing requirements, federal policies and regulations should account for the possibility of an *inadequate* auditing regime leading to a false sense of security about the safety of GPAIS. In recent years, such an effect of the bond rating agencies has been considered a major contributing factor to the 2008 economic collapse. With regard to GPAIS specifically, this outcome may be all the more challenging as research and model evaluations have identified the potential of GPAIS' outputs to deceive or manipulate users.[20]

[18] NIST SP 800-152.
[19] Schuett, 2023 [ibid].
[20] Lin, Hilton, and Evans, 2022. "TruthfulQA: Measuring How Models Mimic Human Falsehoods."; Roff, 2020. "AI Deception: When Your Artificial Intelligence Learns to Lie."; ARC Evals, 2023. "Update on ARC's recent eval efforts."; Fornaciari et al., 2021. "BERTective: Language Models and Contextual Information for Deception Detection."

In addition to an effective assessment regime, contestability should also be promoted and operationalized through human-centered mechanisms for incident reporting, feedback, request for information, and redress for harms.

Contestability mechanisms allow affected stakeholders to challenge decisions and outputs influenced by AI systems. This could mean clarifying a system's output, the data it used, the way it was trained, or other aspects of its operations, and creating space for harm redress, scrutiny, and dialogue, which can enhance the AI system's transparency and foster trust. The relevance of those dimensions to trustworthy AI is recognized in the Blueprint for an AI Bill of Rights' principle of notice and explanation.[21] Contestability can be incorporated into regulatory requirements by requiring operators to provide: (i) notice when people are interacting with or impacted by decisions influenced by generative AI systems or GPAIS, (ii) explanation of how such decisions have been made, and (iii) clear information about and access to appeal procedures.

A salient challenge with generative AI systems and GPAIS is that developers currently lack the technical means of determining with certainty the factors leading to their outputs, and thus, deployers lack means to provide downstream users and impacted persons with an explanation of how a decision was made. End-users could, however, be supplied information about the process and assessments of accuracy and trustworthiness that the AI system has undergone, for instance, evidence that the system has achieved a reasonable performance on a domain-specific benchmark. Most importantly, persons impacted by a decision-making process that involves a generative AI system or GPAIS should be granted the possibility to have the decision reviewed and, if deemed necessary, revised. Contestability is one tool within a broader range of feedback mechanisms and must be coupled with complementary accountability mechanisms to inform people about AI systems' compliance with trustworthy AI standards.

By promoting contestability, NTIA can help ensure that GPAIS-influenced decisions in areas under its purview can be questioned and, if necessary, appealed. Overall, such an approach can support democratic values of transparency and accountability. By encouraging AI developers and deployers to implement robust feedback systems, NTIA can ensure that the voices of all stakeholders, including

[21] The White House, 2022. "Blueprint for an AI Bill of Rights."

traditionally marginalized groups, are heard in shaping AI systems. This can contribute to a more inclusive and equitable digital economy, ensuring that AI systems serve the needs of all users and do not inadvertently reinforce existing inequalities.

Question 14: Which non-U.S. or U.S. (federal, state, or local) laws and regulations already requiring an AI audit, assessment, or other accountability mechanism are most useful and why? Which are least useful and why?

At TFS, we have worked closely with a wide range of multilateral institutions, including the European Union, the OECD, UNESCO, the Global Partnership on AI (GPAI), and the Council of Europe, to develop governance frameworks that promote accountability. For example, members of our organization worked within OECD Working Groups to develop the OECD AI Principles[22] and contributed to their Framework for the Classification of AI Systems.[23] Recently, TFS has advocated for the inclusion of a special governance regime within the EU AI Act to address general-purpose AI systems,[24] and we are also conducting research on an effective enforcement regime. Based on this experience, we believe NTIA should draw from the following governance frameworks.

The OECD Principles on Trustworthy AI can serve as a useful instrument for operationalizing accountability mechanisms while promoting international coordination on safety standards.

The United States adopted the OECD Recommendation on Artificial Intelligence, the first set of intergovernmental principles for trustworthy AI, in May 2019.[25] These principles subsequently served as the basis for US Executive Order 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government."[26]

---

[22] OECD AI Principles.
[23] OECD Framework for the Classification of AI Systems.
[24] The Future Society, 2022. "Memo Fantastic Beasts and How to Tame Them."
[25] U.S. Department of State, n.a. "Artificial Intelligence (AI)."
[26] Executive Office of the President, 2020. Executive Order 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government."

The OECD Principles are useful because they provide clear guidelines for how accountability should be operationalized. The Principles promote a human-centered approach and uphold our values such as respect for the rule of law, human rights, and diversity. They highlight the need for accountability instruments that ensure robustness, security, and safety throughout an AI system's lifecycle, and make clear that transparency is non-negotiable. Lastly, the principles suggest the uptake of contestability mechanisms, to reinforce accountability and public trust in AI technology.

Having adopted these principles, the U.S.—through its agencies, including NTIA—should incorporate them in their policies and regulatory practices.[27] Doing so will also promote international coordination and collaboration on issues and risks related to AI governance.

The EU AI Act can serve as a useful instrument for emphasizing strict internal and third-party assessment requirements, and a robust liability regime.

Deliberations on the draft EU AI Act have revealed noteworthy observations with regard to assurance of accountability, namely, the need for both internal and third-party auditing, for qualified third-party auditors, and for joint liability between developers and auditors.

First, deliberations on the EU AI Act have indicated that auditing requirements should involve combinations of internal and third-party auditing, such as external red-teams, to reflect the risk profile of AI systems. This combination reflects the multifaceted nature of AI-related risks. Internal and third-party auditing processes serve complementary roles.[28] Internal auditing, informed by deep knowledge of the AI system's design and operation, can spot issues and risks that might elude external auditors. Conversely, third-party auditing provides an objective assessment free from potential internal biases, thereby catching oversights that internal auditors might miss. Consequently, incorporating both types of auditing in AI systems significantly improves the chances of identifying and mitigating potential risks. Additionally, this dual approach could promote the advancement of

[27] Cohen et al., 2022. "A Manifesto on Enforcing Law in the Age of "Artificial Intelligence"," Recommendation 1. The Athens Roundtable on Artificial Intelligence and the Rule of Law.
[28] European Parliament, 2023. Draft Compromise Amendments on the Draft Report of the EU AI Act, May 9th.

AI auditing as a whole, improving its practices and standards, and ultimately enhancing the safety, reliability, and transparency of AI systems. Likewise, NTIA should consider a two-pronged auditing approach that has different requirements for internal and third-party assessments.

Second, deliberations on the EU AI Act have highlighted the need for third-party audits to be performed by professionals with demonstrated expertise in the technology they are assessing, e.g. GPAIS. AI systems demand a high level of expertise from external auditors to be able to effectively and rigorously evaluate them. The large-scale demand seems unlikely to be met immediately, and thus, this is likely to pose challenges to the enforcement of law. The lack of expertise is especially dire for GPAIS, where talent is scarce and Big Tech labs invest significant amounts of resources in recruitment. The current draft of the EU AI Act requires that third-party auditors demonstrate their independence, competence, absence of conflicts of interests, and minimum cybersecurity requirements. Similarly, NTIA should require that third-party auditors meet such requirements.

Third, an auditing scheme should require that liabilities are shared between developers, deployers, and third-party auditors. Whereas AI auditing is at a relatively nascent state, AI systems continuously demonstrate new capabilities, many of which are hazardous. Transferring absolute liability to third-party auditors would erroneously presuppose their capability to audit for novel risks. It may even incentivize developers and deployers to take a light approach to internal auditing and risk mitigation, on the assumption that third party auditors would shoulder liability for incidents. Shared liability between developers, deployers, and auditors encourages all involved parties to maintain high standards of diligence, enhances effective risk management, and fosters a culture of accountability in AI development and deployment. We suggest that the federal policies and regulations draw from the EU AI Act experience to build a liability regime appropriate for the US, in order to ensure AI developers and deployers across the value chain are liable for resulting harms to individuals, property, communities, and society.[29]

[29] Future of LIfe Institute, 2023. "Policymaking in the Pause."

THE
FUTURE
SOCIETY

III. Apply a horizontal, cross-sectoral approach to policies and regulations for general-purpose AI systems (GPAIS).

Question 30: What role should government policy have, if any, in the AI accountability ecosystem? (a.) Should AI accountability policies and/or regulation be sectoral or horizontal, or some combination of the two?

Overall, to foster a more resilient AI accountability ecosystem, *federal policy and regulation should advance both outcomes-based and process-based requirements for AI developers and deployers.* Specifically, an outcomes-based approach, drawing on the Blueprint for an AI Bill of Rights and NIST's Risk Management Framework, could favor responsible innovation and provide a leeway for companies to adjust their practices according to their contexts. Outcomes-based requirements allow for companies to apply different technologies to testing and assessing their AI systems' robustness and trustworthiness, instead of requiring a specific technique to be applied. This approach avoids distorting the innovation in potential technical and institutional safeguards that GPAIS providers will likely develop and leaves incentives for the development of better techniques.

Broadly speaking, until now, the federal approach to AI risk management has been sector-specific.[30] This approach may be sufficient for more narrow AI systems since their specific use cases and associated risks can be narrowly addressed. However, it is not sufficient for GPAIS which, due to their cross-sectoral nature and opaqueness, require a transversal, horizontal approach.

We are concerned that a lack of horizontal regulation in the US could perpetuate a regulatory vacuum and "race-to-the-bottom" dynamics among GPAIS developers, as they increasingly develop technologies that can pose risks to public health, safety, and welfare in an unregulated environment.

Additionally, a sectoral approach to regulation could lead to multiple regulations and mounting compliance requirements, which would create a disproportionately

[30] Engler, 2023. "The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment." Brookings Institute.

heavy burden on small AI innovators. It is noteworthy that SMEs represent over 99% of U.S. enterprises.[31]

As the market for AI systems becomes dominated by a small number of big tech companies,[32] federal policy and regulation has the responsibility to promote competition by enforcing effective horizontal regulation that not only targets deployment, but also includes the development stage of GPAIS.

Such an approach doesn't require stifling innovation; on the contrary, holistic federal policies and regulations have the potential to promote responsible innovation and secure U.S. leadership in a global landscape.

We, at The Future Society, appreciate this opportunity to comment on these issues and NTIA's efforts toward an effective AI assurance ecosystem. We welcome any further opportunity to provide resources or information to assist in this important effort. If you have any questions regarding these comments and recommendations, please contact Niki Iliadis at niki.iliadis@thefuturesociety.org (cc:info@thefuturesociety.org).

Sincerely,

Niki Iliadis,
Director, AI and the Rule of Law, The Future Society

Amanda Leal,
Associate, The Future Society

Samuel Curtis,
Associate, The Future Society

The Future Society
867 Boylston Street, 5th Floor
Boston, MA, 02116

[31] OECD, 2022. "Financing SMEs and Entrepreneurs 2022: An OECD Scoreboard."
[32] Roose, 2023. "How ChatGPT Kicked Off an A.I. Arms Race." The New York Times.