CIfAI

Cognitive Insights for Artificial Intelligence

Response to the National Telecommunications and Information Administration (NTIA), U.S. Department of Commerce's request for comment on artificial intelligence (AI) system accountability, measures and policies, Docket No. 230407-0093
Submitted by Monica Lopez, PhD and Irene Gonzalez, PhD.
Organization, Cognitive Insights for Artificial Intelligence (CIfAI)
June 6, 2023

On behalf CIfAI, we write in response to the NTIA, U.S. Department of Commerce's request for comment on artificial intelligence (AI) system accountability measures and policies. The multifaceted nature of advancing trustworthy AI is a highly complex one, and we appreciate and support NTIA's call on AI Accountability to establish a robust infrastructure of harm assessment and mitigation and engender trust for all stakeholders.

We at CIfAI provide strategic research-based solutions from a human-centered perspective to ensure the safe and ethical design, development, deployment, and management of AI-enabled autonomous systems across various industries. Our values-based approach is founded on accuracy, consistency, and context-dependency, and supports trusted data across every phase of the AI lifecycle to achieve confident and fair decision making.

CIfAI has reviewed all 34 questions provided in NTIA's request for comment and provide several recommendations (total 12) to a subset of questions (#1, 2, 3, 5, 7, 9, 10, 16, 18, 22, 23, 29) below.

As AI-enabled systems proliferate across a variety of use cases and influence our day-to-day decisions, AI assurance becomes ever more imperative. The need to operationally define 'trustworthy AI' and thus support AI assurance has led to many proposals across entities and sectors, with the noteworthy development of principles since 2019 by the Organization of Economic Co-operation and Development (OECD) and their subsequent full or partial adoption thereof by many into implementable tools. However, as principles are integrated within emerging accountability mechanisms such as AI assessments and audits, a multitude of new (non-exhaustive) questions arise within established focus areas even as criteria use current and emerging legal standards as a baseline. The following are ongoing research questions:

1. Bias and discrimination: Cutoff points and metrics.

    a. What other factors that could lead to bias and discrimination have not yet been considered?
    b. What constitutes dynamic assessment and mitigation of harm and assurance thereof?
    c. When is a system sufficiently non-biased and non-discriminatory? What is an acceptable boundary of statistical bias and resulting performance?
2. Effectiveness and validity: Unintended or unforeseen outcomes and/or use.
    a. Are simulations during testing and evaluation sufficiently realistic?
    b. What other non-technical factors have been considered and consulted upon within the contexts of both product development and overall organizational governance and purpose around AI technology development and/or use?
    c. How robust is the system in predicting attack and compromise?
3. Data protection and privacy: Values tradeoff and data quantity and quality.
    a. What is the risk level of available and accessible data?
    b. Are available and accessible data sufficient and complete?
    c. Who has access to the data and when? What data sharing agreements are in place?
4. Transparency and explainability: Stakeholder benefit.
    a. What is the best method of explainable AI?
    b. Explainable AI for what purpose and for whom?
    c. When is explainable AI most pertinent?

We advance the proposal that these (and other) emerging questions must be determined in a collective-driven and consensus-bound manner. Moreover, we advocate determinations should be founded in mitigating risks to humans and thus achieved through a human-centered approach and supported via robust long-term empirical work between regulatory agencies, industry peers and academic partners. The above necessitates answering to significantly advance the utility and standardization of AI assurance measures, including the establishment of an oversight body that approves audit criteria and oversees certifying bodies.

Responses to Specific Questions

Question 1. What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?

Principally, the purpose of AI accountability mechanisms is to provide a robust infrastructure in which to develop, assess, mitigate and assure an algorithm's legality, ethics and safety. Any AI system deployed, therefore, is evaluated for its purpose and objectives, its benefits, and its risks. AI accountability allows for a system to be explained, subserving a user's right to an explanation and control over their personal information (e.g. the General Data Protection Regulation (GDPR)'s "meaningful information about the logic involved" (Article 13); the California Consumer Privacy Act (CCPA)'s disclosure, use limitations, deletion and correction of personal information); allowing for designers and developers to enhance system robustness; enabling the prevention of bias, unfairness, discrimination, and the like; and increasing overall technological acceptance as users maintain awareness for whether an AI-enabled system's decisions are properly accounted.

Recommendation #1: Given the emerging nature of annual activities reporting (e.g. European Commission's Digital Services Act, Article 44), the five principles of fair AI practices proposed in the OSTP's Blueprint for an AI Bill of Rights (safe and effective principles; algorithmic discrimination protections (fairness and equity); data privacy; notice and explanation (transparency); human alternatives, consideration, and fallback (accountability)), and the questions described above, we offer as a starting point for multi-stakeholder discussion a framework for algorithm audits that captures five key risk verticals: Robustness, Bias, Privacy, Explainability, and Efficacy (as derived from the typology developed by Koshiyama et al. (2022)).

Question 2. Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

The goal of an AI audit should be to improve a user's confidence in a system's capacity. While certifications function as public-facing documentation on, for example, a system's level of reliability and thus safety, internal assessments help to improve a system at the R&D level, directly guiding better decision-making and best practices across the conceptualization, design, development, and management and monitoring of a system. External audits offer yet another level of system assurance through the process of independent and impartial system evaluation whereby an auditor with no conflict of interest can assess the system's reliability and in turn identify otherwise unidentified errors, inconsistencies and/or vulnerabilities. As such, internal assessments of performance according to clearly delineated criteria are necessary for internal purposes as much as for providing the documentation trail (e.g. logs, databases, registers) of evidence of system performance for external independent and impartial auditing.

Recommendation #2: Internal assessments, external audits and certifications are all necessary components for AI assurance and should be standardized for maximum efficiency. Audit criteria should be empirically determined, collectively approved and overseen by an oversight body.

Question 3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?

All the goals listed are very complex and interconnected given the multifaceted nature of AI systems. As such, an interdisciplinary team of experts across various domains (e.g. computer science, cognitive science, psychology, anthropology, philosophy, business, law, government) is paramount to best identifying a prioritization of goals according to use-case context, and then integrating the multitude of factors and diverse perspectives particular to each goal. There is no one-size-fits-all solution.

Recommendation #3: A body of interdisciplinary experts needs to collectively determine best practices, standards and regulations to ensure inclusion of a diverse range of interests and policy needs. This body should be composed of stakeholders beyond, for example, the big technology players of the private sector and large international NGOs; such stakeholders should include smaller technology companies and local civil society organizations given their frontline work with users.

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?

AI accountability mechanisms can inform users about the quality of already included system guardrails, as well as inform users about the weaknesses of such guardrails. Due to the large-scale use of large language models (LLMs) and other generative AI systems, AI accountability mechanisms become critical to highlighting the level of risk of, for example, nonfactual, inaccurate and/or harmful outputs; IP infringement; disclosure of sensitive data; and adversarial attacks.

Recommendation #4: LLMs and other generative AI systems necessitate governing and therefore AI accountability mechanisms to enable a trustworthy space for generative AI.

Question 7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

No, accountability mechanisms are paramount to understanding system harm assessment and mitigation and therefore critical for eliciting user confidence in the AI system. Scientific innovation and technological advancement result from challenges and opportunities.

Recommendation #5: Accountability mechanisms are vital at this critical moment in time and offer an opportunity to collectively evaluate and transform the development of AI for the benefit of humanity through the implementation of resilient guardrails within AI.

Question 9. What AI accountability mechanisms are currently being used? Are the accountability frameworks of certain sectors, industries, or market participants especially mature as compared to others? Which industry, civil society, or governmental accountability instruments, guidelines, or policies are most appropriate for implementation and operationalization at scale in the United States? Who are the people currently doing AI accountability work?

CIfAI suggests the use of independent and impartial AI audits using a service platform for AI governance, risk management, and regulatory compliance. Audit criteria have been developed via multistakeholder collaboration and empirical work.

Recommendation #6: As an established entity with clients across the U.S. and the world, auditing platforms should be open to working with NTIA and others to further develop sector specific accountability frameworks as prescribed by law and emerging regulatory regimes.

Question 10. What are the best definitions of terms frequently used in accountability policies, such as fair, safe, effective, transparent, and trustworthy? Where can terms have the same meanings across sectors and jurisdictions? Where do terms necessarily have different meanings depending on the jurisdiction, sector, or use case?

It is important to highlight that the definitions of AI and related topics are not yet fully agreed upon (cf. Lost in Transl[A]t[I]on: Differing Definitions of AI). As such, the following listed are broad definitions:

- Fair: The system is equal and equitable in the treatment of individuals given their protected characteristics.
- Safe: The system does not cause harm, or at the very least has a robust mitigation strategy in place to prevent harm.
- Effective: The system works as expected with a range of inputs and in a variety of situations.
- Transparent: The system's decision-making process is understood by all stakeholders.
- Trustworthy: The system can be deemed trustworthy if the following impactful criteria are satisfied: performance and robustness, bias and discrimination, interpretability and explainability, and algorithm privacy.

Recommendation #7: The above definitions should be consistent across all AI system domains to engender trust from all and protect all from harm no matter the AI system developed. However, agreement on definitions will depend on balancing the technical with human-based definitions (i.e., analogies to human intelligence and capabilities). Particularly, system capability-based definitions that encompass both classical algorithms and statistical techniques and modern complex systems will be paramount to supporting broad enough scope and legal precision. Additionally, focusing on the impacts of AI systems, i.e., systemic risks to humans, supports the evaluation of system design and functioning as it relates across domains and underscores any differences across user groups, incidents, and regional contexts.

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."

Given the dynamic nature of AI systems and dependency on data to learn and new data –real or synthetic– to improve, internal assessments of performance require continuous updates and therefore renewal of audits. This is a unique challenge of adaptive systems whereby a system deemed compliant at one point may not be compliant later. Annual auditing requirements may be insufficient for certain systems and their contexts.

Recommendation #8: System improvements can be small or large so appropriate recertification depends on the identified magnitude of risk at present. Documentation of continuous internal assessments becomes critical to identifying a system's current level of risk.

Question 18. Should AI systems be released with quality assurance certifications, especially if they are higher risk?

Yes. Users deserve the right to know:

1. That the AI-enabled product or service they are consuming is reasonably safe for use.
2. Any identified risks the AI-enabled product or service may incur because of its use.
3. Safety measures that have been put in place to assure confidence of use.

The above assumes that a cutoff point of quality has been determined prior to market release. Moreover, the above supports transparency if an AI-enabled product or service fails to be certified, indicating that it is not assured or trustworthy in its present iteration.

Recommendation #9: Quality assurance certifications should be mandated and presented in a clear, concise and non-jargon heavy manner.

Question 22. How should the accountability process address data quality and data voids of different kinds?

Due to a host of limitations on data availability (e.g., privacy, security, time frame of data collection, unintended errors), the absence of certain data points is inevitable. Filling in data gaps is possible only with the right amount of information.

Recommendation #10: There should be concise documentation on the what, when, where and why of data. This will allow for gaps to be filled for subsequent audits.

Question 23. How should AI accountability "products" (e.g., audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?

The following should be standardized, made transparent, and communicated concisely to the public:

1. Audit certification rules.
2. Outcome (e.g., success, failure) of compliance with audits.

Recommendation #11: The way privacy labels are emerging to provide educational, convenient and readable information on the security and privacy components of a technology, certification labels could be developed to provide key information on the Robustness, Bias, Privacy, Explainability, and Efficacy of an AI system. This would serve multiple goals:

i. to re-evaluate AI product and service developers' own organizational practices on AI assurance;
ii. to support healthy competition between AI product and service developers;
iii. to improve the presentation of audited information to the consumer; and
iv. to empower consumers to smartly choose between products that best align with their values.

Question 29. How does the dearth of measurable standards or benchmarks impact the uptake of audits and assessments?

The dearth of measurable standards or benchmarks underscores the need to invest resources in more research on critical areas of trustworthiness such as transparency, explainability, and data management across the AI lifecycle.

Recommendation #12: The establishment of a collaborative and robust relationship between government, industry and academia to innovate and test out new methods. A regulatory sandbox-type arrangement would be pro-innovation.