Computing Community Consortium's Response to RFC on AI Accountability Policy Request for Comment

June 12, 2023

Written by: *Nadya Bliss (Arizona State University), David Danks (University of California, San Diego), Maria Gini (University of Minnesota), Jamie Gorman (Arizona State University), William Gropp (University of Illinois), Madeline Hunter (Computing Community Consortium), Odest Chadwick Jenkins (University of Michigan), David Jensen (University of Massachusetts Amherst), Daniel Lopresti (Lehigh University), Bart Selman (Cornell University), Ufuk Topcu (University of Texas at Austin), Tammy Toscos (Parkview Health), and Pamela Wisniewski (Vanderbilt University).*

Introduction: Scoping AI for Accountability

The term Artificial Intelligence (AI) encompasses a very broad and vastly complex set of technologies. As research and development of AI rapidly proliferates, and these systems are increasingly ubiquitous in their impacts on everyday life, there is an urgent need to recognize, control and mitigate the adverse consequences they can have on citizens. We commend the Federal government for its efforts, such as this request for comment, to obtain feedback from the public and relevant communities to better understand the potential issues AI systems create and explore possible solutions.

AI systems are incredibly complicated and diverse systems that make it very difficult to propose or even think about mechanisms by which to hold them accountable. We must ask ourselves what we specifically want to audit, which entities should be empowered as auditors, and whether there are even entities that are capable of such feats. Through this response, we aim to address common themes and lenses that will help in incorporating differences between different approaches to audit and accountability of AI systems, as well as how to think about these questions in a more nuanced, specific approach.

Before turning to specific aspects of the request for comments, we note two specific points that were not necessarily clear in the call, but which are important to ensure shared community understanding. First, the call seems to use the terms recourse, accountability, transparency, and audit interchangeably, using accountability as a catch-all term. We suggest that it would be helpful to articulate that (and how) these are distinct notions with important implications for one another. For example, if we can audit the system's data or other system characteristics to ensure a responsible system, then issues of accountability would potentially be less pressing. Second, given the plethora of entities encompassed under the "AI Systems" umbrella, it could be helpful to define a specific scope of what is meant by AI when we are considering accountability. For instance, there could be a prerequisite category of questions such as, "What should or should not be considered AI for the purpose of holding it accountable?" or "How is AI significantly different from other

systems such that we need to develop new mechanisms for accountability?". We turn now to more specific responses.

AI Methods

Different levels of audit

Objectives and the general scope of AI accountability will vary across systems and particular fields. In order to get a comprehensive understanding of what can and should be audited, we must take into account different levels of audit. There are generally four levels to keep in mind:

1. Quality of data independent of models;
2. Outputs of the model in controlled testing;
3. Performance of the system in real-world deployment;
4. Impacts and effects of the system on people.

Each level of audit contains very different issues and needs that are a necessary dimension in framing these questions and solutions relating to AI accountability. A particularly timely example would be ChatGPT. With such a multi-purpose tool we have to think about which level of auditing would be most effective in ensuring accountability – auditing the actual outcome or answer that ChatGPT produces, what the user does with the answer ChatGPT provided, ChatGPT's performance in "lab" settings, or possibly even auditing the data used to train ChatGPT. Diving deeper into the first dimension – accountability in datasets – we have little to no actionable understanding of what data is being used by these systems. Even if information were shared about the data, there would still be a lack of understanding regarding how to use that knowledge to audit. We need to keep these levels of audit in mind when considering the scope and objectives of AI accountability. There are places that conflate audit assessment with accountability. The AI system can be audited for performance output, in instances when we should really be looking at the harm and effects.

Accountability methods beyond audits

There are other possible "guardrails" which may help suppress the growth of the undesirable outputs and may even prevent some of them beyond audit. The decision of which accountability method to apply is dependent upon the field that the AI system operates within and the contexts in and use cases for which the system is to operate. For example, if we are seeking to certify an autonomous aircraft, we make assumptions on traffic controls, where we would fly it, and other real world factors. It is vital that all methods and objectives of AI accountability bear in mind the context and use case. Furthermore, one cannot use a single and unified mechanism in every field. There is a need for increasing the diversity of methods for accountability and establishing practices for adapting the resulting toolset to the needs and challenges of the field.

An additional factor to think about is retraining AI systems as they are going to be retrained very frequently because they are based on machine learning. It is already

challenging to think about certifying complex systems, and integrating AI-supported functionality will increase system complexity, once. The idea of recertifying the system every time a new version is trained is near impossible. In these cases, there is a need for formalisms and practices that will support certification throughout the lifetime of the system.

Accountability Subjects

There are different approaches to the timing of audits or assessments, again depending on the context and level of accountability that you are looking at. One means of practicing AI accountability, briefly mentioned above, is validating the code and data at level one, which is best done at the beginning of the process before the technology is implemented. The bias in outcomes comes from the datasets that the system is trained on. Making sure the data are comprehensive, accurate and adequate representative could prevent potential harms from biased decision making from ever happening. When looking at the outcome directly, after the fact, we are missing the input layer. If we can make things right at the beginning, in the data sets that we use, then we don't have to consider who to hold accountable, especially in low stakes domains. This is especially important in high stakes, decision making instances, when what really matters is not making the mistake in the first place. We must ask ourselves how and why we are attempting to regulate the systems in order to reach accurate assessments on the best means and time frame to conduct audits and assessments.

Alternatively, there are instances when the centrality of the issue is not on the system input, but the output and the effects these outputs have on a user. For example, again going back to ChatGPT, although the inputs are important to the outcome, especially in cases of disinformation, the crux of the issue stems from how the user uses the tool and what they do with the information they receive. For this type of instance cases is, if not impossible, extremely difficult to ensure accountability through validating datasets. A more comprehensive approach for this sort of system could be a more dynamic, on-going evaluation. For example, if we are auditing a human, we are using what we know about the person and their past behavior. In the same way, with systems, we could seek auditability by recording a system's behavior overtime to prove it will not do anything bad and flag potential harmful behaviors to look out for.

Existing Resources and Models

In some cases, we already measure outcomes. We should leverage what already exists in the government to do these measurements and develop methods to figure out how to do this. Keeping in mind the nuances and distinctions across fields, one way to do this would be compiling enumerations of best practices by experts and current models to use as a foundation to adapt to new situations.

One potential example model for guiding auditing decisions that is based on interaction with human subjects in human subjects research is the classification of human-AI interactions in terms of their foreseeable risk to humans, beyond those of everyday risk (the Common Rule; HHR 25 CFR 46). This classification of human subjects

research (exempt – no risk; expedited – minimal risk; full review – risks beyond the level of minimal risk but benefits outweigh risks) serve to determine the level of scrutiny and approval a human subjects research protocol must undergo to be approved. It is likely that a similar ranking system could be developed for human-AI interaction to guide the level of audit and evaluation that an AI technology undergoes with respect to its impact and risk for its users.

Another potential model for guiding auditing decisions that is based on interacting with human subjects for research purposes are the 1978's Belmont Report's principles of (1) Respect for Persons, (2) Beneficence, and (3) Justice. All interactions with human subjects for the purposes of research must meet these standards. With respect to interacting with AI technologies, Respect for Persons would entail the consideration of a person's autonomy in making the decision to use the technology – for example, do they have adequate information of how the system works, its competence envelope, its potential benefits, and any risks involved in using the technology. A second ethical consideration from this principle is consideration of a person's autonomy in using the technology – for example, people with diminished autonomy (e.g., children; people with mental disabilities) merit additional protections before using the technology. Another aspect of this principle is coercion. Are people adequately informed about their options for not using the technology or alternative technologies that could be used.

In this context, consideration must be given to long-term benefits and risks that human interaction with AI technologies could involve through forethought. In this context, Justice would entail a consideration of who receives the benefits of the technology and who bears its burdens. For example, the Tuskegee syphilis study of the 1940's involved rural black men to study the course of the disease; however, although they bore all of the risks, the men accrued no benefits of the findings, constituting an injustice. AI technologies, particularly those with potential economic, social, or health benefits should be developed and shared with this principle in mind. Broadly, AI technologies with either risks or benefits should be justified in the sense that they are not biased with respect to who contributes to their development or who can access them. It is likely that AI technologies evaluated using these or similar standards will promote safety, inclusion, equity, and transparency in human-AI interactions.

Clearly, there are inherent differences between fields that make a single model impossible to apply across all AI systems, but these existing models, guidelines and tested best practices can be used as a foundation to establish new models adapting to different fields and use cases.

Accountability Inputs, Transparency, and Barriers

There are many frameworks in computing that can inform a sound approach to creating accountability in the design, development and maintenance of AI systems. One example is human-centered AI, which is built on the foundation of convening interdisciplinary and inclusive conversations around the design of software systems. Compelling discussions around ethical concerns and biased algorithms could be encouraged if there is a

requirement for companies to include this type of approach in the development and evaluation of their systems.

Organizations that create and use AI systems should be required to be more transparent about the flaws in their models. While there is a need to protect intellectual property, the lack of external validation creates an imbalance that we would not allow in other industries that potentially could result in loss of life. For example, prescription medications are heavily regulated by the FDA to protect individuals from injury or death. As a result, consumers are informed of all possible side effects and the mechanism behind how the drug interacts in the body. Why aren't the same safeguards in place for predictive models that have been built by health IT companies to identify critical, life-threatening medical situations (e.g., sepsis for patients that are hospitalized, chatbots targeting individuals with a mental health crisis).

In the best case scenario where organizations allow for external validation of models, there may be additional barriers related to data access. Using healthcare again as an example, there are very specific protections in place protecting patient privacy. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is a federal law that prevents organizations from sharing patient information without their consent. For some AI systems (e.g., a sepsis predictive model) it would require an actual dataset – including all of the flaws seen in electronic health record data - to validate the model. Thus, there would need to be a neutral third party organization that would be given permission to access data and algorithms in order to create accountability in AI systems.

As with all types of oversight systems, there will be a significant cost. There should be a thoughtful approach to building this oversight in order to avoid costs being passed down to the consumer, as we have seen occur with prescription medications in this country. One possibility would be to incentivize organizations to be fully transparent with algorithms and flaws they have detected in their own testing. The development of such oversight would likely need to be industry specific given the different challenges that are experienced and expertise required to overcome unique barriers (e.g., healthcare vs. transportation industries).

We also encourage taking a look at the CCC's past workshop report on Assured Autonomy, which discusses relevant issues and ideas.