June 12, 2023

National Telecommunications and Information Authority
U.S. Department of Commerce
Mr. Travis Hall
1401 Constitution Avenue NW, Room 4725,
Washington, DC 20230

Re: AI Accountability Policy Request for Comment

Dear Mr. Hall,

On behalf of the Center for Data Innovation (datainnovation.org), I am pleased to submit this response to the National Telecommunications and Information Authority's (NTIA) request for comments on AI accountability measures and policies.[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

Please find our responses to the following questions in the document below.

[1] "AI Accountability Policy Request for Comment," Federal Register, April 13, 2023, https://www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment.

Sincerely,

Hodan Omaar
Senior Policy Analyst
Center for Data Innovation
homaar@datainnovation.org

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?

There is a trade-off between explainability and accuracy. Explainable AI systems are those that can articulate the rationale for a given result to a query. Explanations can help users make sense of the output of algorithms and may be useful in certain contexts, such as to discover how an algorithm works. Explanations can reveal whether an algorithmic model correctly makes decisions based on reasonable criteria rather than random artifacts from the training data or small perturbations in the input data. In certain scenarios, some users may also be more likely to trust explainable AI systems.

However, one cannot maximize explainability without a loss of system accuracy as researchers from Stanford and the University of Chicago explain in a 2021 paper called "Unpacking the Black Box: Regulating Algorithmic Decisions." They write, "we can restrict algorithms to produce prediction functions that are simple enough to be fully transparent, e.g., a ten variable logit model, but sacrifice the predictive performance that complex algorithms provide. Alternatively, we can allow complex algorithms but sacrifice some of their ability to understand the model and detect actions that arise from incentive misalignment."[2]

When it comes to increasing user trust, accuracy can be a more decisive factor than explainability. A 2019 study led by researchers from the Leibniz Institute of the Social Sciences in Germany measured how much trust 327 participants had in systems that detect offensive language in tweets with varying degrees of accuracy.[3] They found that, in general, the more accurate a system was, the greater trust users had in the system. But the effect of explanation accuracy (the probability an explanation is true) on trust was more complex. In highly accurate systems, for example, any explanation, whether the explanation was accurate or not, decreased how much users trusted the system. This is because when individuals learn new information, they have to reconcile it with their existing understanding. When dealing with highly accurate systems, explanations that provide new information or a new way of understanding make users question their mental model, leading to decreases in trust. But in systems with medium levels of accurate results, a highly accurate explanation had no impact on user trust and a less accurate explanation decreased trust.

In order to be efficient, audit tools have to find the optimal balance between accuracy and explainability. Indeed, the researchers from Stanford and the Univeristy of Chicago note that the

---

[2] Laura Blattner, Scott Nelson, Jann Spiess, "Unpacking the Black Box: Regulating Algorithmic Decisions," EC '22: Proceedings of the 23rd ACM Conference on Economics and Computation, (July 2022), https://doi.org/10.1145/3490486.3538379.

[3] Andrea Papenmeier et al, "How model accuracy and explanation fidelity influence user trust in AI" (July 2019), https://arxiv.org/pdf/1907.12652.pdf.

"optimal algorithmic audit is a targeted explainer that requests information not about what drives the average prediction but instead about what drives particular types of mis-prediction." That is, algorithmic audits should not focus on every algorithmic decision being explainable, but rather focus on inspecting parts of a model that are most likely to create harmful model distortions.

6. The application of accountability measures (whether voluntary or regulatory) is more straightforward for some trustworthy AI goals than for others. Are there any trustworthy AI goals that are not amenable to requirements or standards? How should accountability policies, whether governmental or non-governmental, treat these differences?

As professors Ellen Goodman and Julie Trehu rightly explain in their 2022 report *AI Audit-Washing and Accountability*, AI accountability measures can serve as means to various ends. These ends range from confirming compliance with narrow legal standards to enquiring about broader ethical commitments, but as the authors note, the functional objectives of audits fall into five broad categories:

1. Fairness; meaning audits check whether systems are biased against individuals or groups as it relates to demographic characteristics;
2. Interpretability and explainability, meaning audits check whether systems make decisions or recommendations that users and developers can understand;
3. Due process and redress, meaning audits check whether systems provide users with adequate opportunities to challenge decisions or suggestions;
4. Privacy, meaning audits check whether the AI systems protect individuals' privacy rights and adhere to relevant laws, regulations, and best practices related to data privacy; or
5. Robustness and security, meaning audits check that systems are operating as intended and performing consistently and accurately under various conditions, including adversarial attacks.[4]

Some of these goals are more amenable to requirements and standards than others because they can more easily be translated into objective, concrete metrics. For instance, to ensure AI systems are robust and secure, one can employ audits to check how prevalent algorithmic errors are. There are various types of error-analysis techniques to check for algorithmic error, including manual review, variance analysis (which involves analyzing discrepancies between actual and planned behavior), and bias analysis (which provides quantitative estimates of when, where, and why systematic errors occur, as well as the scope of these errors).

[4] Ellen P. Goodman and Julia Trehu, "AI Audit-Washing and Accountability," (German Marshall Fund, November 2022), https://www.gmfus.org/news/ai-audit-washing-and-accountability.

But other goals, such as fairness, are subjective and cannot be reduced to fixed functions. To see why, consider two housing authorities that are using AI systems to achieve the same policy goal: allocating financial support through housing assistance programs. One housing authority uses a system whose objective function is to minimize the total number of families that experience eviction. The other housing authority uses a system whose objective function is to first provide the family who is most likely to be evicted with as much assistance as possible, then move on to the next, until the budget is exhausted. Assume both systems are completely error-free. As researchers from Harvard, Cornell, and Princeton University show in a 2020 paper, the objective function chosen in this sort of scenario can target very different groups of people.[5] Even if the systems are error-free and working completely as intended, they would have disparate outcomes because they formalize the problem in different ways. It does not make sense to come up with one fixed notion of fairness or audit based on that definition. Rather, mechanisms for accountability should be wholistic, considering the effect of any particular decision system (whether algorithmic or human) on inequality as a whole.

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?

An assumption underlying many calls for algorithmic accountability is that individual companies have the ability and responsibility to wholly correct for algorithmic harms, and that if every company ensured their own actions minimized and prevented algorithmic harms, overall welfare would be maximized. However, recent research suggests that in some contexts, there are factors affecting harm that are outside any individual company's control and in fact, rushing to impose accountability measures on individual companies might have an overall negative impact on fairness.

This outcome is a potential consequence of algorithmic monoculture, which is when multiple decision-makers (or firms) deploy the same systems, or systems that share components such as datasets and models.[6] For instance, imagine multiple firms using the same algorithmic model to screen resumes of job candidates. This scenario is close to the real-world context, more than 700 companies including over 30 percent of Fortune 100 companies rely on a single vendor's tools for resume screening. What recent research suggests is that even if the algorithmic screening tool is

[5] Rediet Abebe, Jon Kleinberg, & S. Matthew Weinberg, "Subsidy Allocations in the Presence of Income Shocks," Proceedings of the AAAI Conference on Artificial Intelligence (2020): 34(05), 7032-7039, https://doi.org/10.1609/aaai.v34i05.6188.
[6] Jon Kleinberg and Manish Raghavan, "Algorithmic monoculture and social welfare," PNAS (2021), https://doi.org/10.1073/pnas.2018340118.

more accurate than human evaluators and less error-prone than other tools on the market, accuracy may become worse when multiple firms use the same ones.[7]

This counterintuitive result is somewhat like the Braess paradox, an observation German mathematician Dietrich Braess made that illustrates how individual entities choosing their most rational option can lead to lower overall welfare when collective interaction is involved. The paradox states that when one adds a new road to a road network it can slow down overall traffic flow rather than speeding it up because individual drivers act selfishly. Drivers want to get from point A to point B in the fastest time, so if the new route is the most efficient way to get to their destination, all drivers will choose to take it. Choosing the new route would be optimal if only one driver did it, but if they all do it, the route becomes suboptimal. Similarly, companies seeking to fill job vacancies want to choose the best performing hiring tools. But it could be the case that if every company chooses to use the same system, there are more errors overall even though the system is the most accurate one on the market.

The reason this can happen in the hiring context is based on the probabilistic properties of rankings. Rather than diving into the math of it, instead consider that for each firm there is some "true ordering" of best to worst candidates for a job role and when using a system (or humans), a firm is using a ranking that best approximates that true ordering. When two firms use a single system to screen candidates, they rely on a common ranking that is a single approximation. Research suggests that even if a single approximation is more accurate in isolation, it can create more errors overall if multiple entities use it.[8] It is better for multiple entities to use different approximations, even if those approximations are less accurate. The key takeaway is that in the hiring context, independence can be more important than accuracy for reducing errors. Importantly, this may not be the case for all settings. Algorithmic monoculture could be desirable in some settings as the authors themselves posit. It may be the case that in other high-risk areas, multiple decision-makers using a single centralized algorithmic system may reduce errors. In education, for instance, economists have found outcomes have improved as algorithms for school assignment have become more centralized.[9] Perhaps in healthcare, the allocation of scarce resource by different hospitals would be best done if they all used the same algorithmic systems. Perhaps not. We do not know because it has not been studied yet.

---

[7] Ibid.
[8] Ibid.
[9] Atila Abdulkadiroglu, Nikhil Agarwal & Parag A. Pathak, "The Welfare Effects of Coordinated Assignment: Evidence from the NYC HS Match," National Bureau of Economic Research (2015), https://www.nber.org/papers/w21046.

Before rushing to regulate and potentially causing negative unintended consequences, policymakers should investigate how different factors affect desired outcomes such as fairness in different contexts.

15a. Where in the value chain should accountability efforts focus?

As the Center for Data Innovation explored in its 2018 report "How Policymakers Can Foster Algorithmic Accountability," regulators should focus their oversight on operators, the parties responsible for deploying algorithms, rather than developers, because operators make the most important decisions about how their algorithms impact society.[10]

This oversight should be built around algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes. When an algorithm causes harm, regulators should use the principle of algorithmic accountability to evaluate whether the operator can demonstrate that, in deploying the algorithm, the operator was not acting with intent to harm or with negligence, and to determine if an operator acted responsibly in its efforts to minimize harms from the use of its algorithm. This assessment should guide their determination of whether, and to what degree, the algorithm's operator should be sanctioned. Regulators should use a sliding scale of enforcement actions against companies that cause harm through their use of algorithms, with unintentional and harmless actions eliciting little or no penalty while intentional and harmful actions are punished more severely.

Defining algorithmic accountability in this way also gives operators an incentive to protect consumers from harm and the flexibility to manage their regulatory risk exposure without hampering their ability to innovate. This approach would effectively guard against algorithms producing harmful outcomes, without subjecting the public- and private-sector organizations that use the algorithms to overly burdensome regulations that limit the benefits algorithms can offer.

16a. Should AI accountability mechanisms focus narrowly on the technical characteristics of a defined model and relevant data? Or should they feature other aspects of the sociotechnical system, including the system in which the AI is embedded?

It is critical that accountability mechanisms do not narrowly focus on the technical characteristics of a model but rather considers the broader sociatechnical system because even even when algorithms can do good by making existing processes more efficient and equitable for consumers, public

[10] Joshua New and Daniel Castro, "How Policymakers Can Foster Algorithmic Accountability" (Center for Data Innovation, May 2018), http://www2.datainnovation.org/2018-algorithmic-accountability.pdf

backlash and opaque implementations can erode the trust needed for them to achieve impact. The experience of the Boston public school system should serve as a cautionary tale to U.S. regulators for what can happen when this is not the case.

In 2018, the Boston public school system proposed using an algorithmic system to improve school busing in ways that would cut costs by millions of dollars a year, help the environment, and better serve students, teachers, and parents.[11] The district had two aims, the first of which was to cut transportation costs. More than 10 percent of the public school system's budget goes toward busing children to and from school—the district's annual cost per student is the second highest in the United States.[12] The district's second goal was to reconfigure school start times so that high school students could get more sleep, as early school starts for teenagers has been linked to serious health issues such as decreased cognitive ability, increased obesity, depression, and increased traffic accidents. Indeed, the American Academy of Pediatrics recommends that teenagers not start their school day before 8:30 AM, but only about 17 percent of U.S. high schools comply.[13]

Boston public school officials engaged researchers from the Massachusetts Institute of Technology (MIT) to build an algorithm to achieve its twin goals, which they did. The Boston Globe called their solution a "marvel."[14] The algorithm helped the district optimize bus routes, cutting 50 of the 650 school buses used, $5 million off the budget, and 20,000 pounds of carbon emissions each day while also optimizing bell times. Importantly, the algorithm's solution for bell times addressed inequity. In the past, the district manually staggered start and end times, but its approach predominantly provided wealthier and whiter schools with later start times while schools with poorer and minority students disproportionately shouldered earlier times. In contrast, the algorithm's solution distributed advantageous start times equally across major racial groups, while significantly improving them for students in all of those groups. Under the status quo, white students were the only group with a plurality (39 percent) enjoying start times in the desirable 8:00 AM to 9:00 AM window but under the algorithmic-determined schedule, a majority of all students (54 percent) in every ethnic group would have start times in that window.

---

[11] David Scharfenberg, "Computers Can Solve Your Problem. You May Not Like The Answer," The Boston Globe, September 21, 2018, https://apps.bostonglobe.com/ideas/graphics/2018/09/equity-machine/
[12] Ellen P. Goodman, "The Challenge of Equitable Algorithmic Change," The Regulatory Review (2019), https://www.theregreview.org/wp-content/uploads/2019/02/Goodman-The-Challenge-of-EquitableAlgorithmic-Change.pdf.
[13] Centers for Disease Control and Prevention, "Most US middle and high schools start the school day too early," news release, August 6, 2015, https://www.cdc.gov/media/releases/2015/p0806-school-sleep.html.
[14] David Scharfenberg, "Computers Can Solve Your Problem. You May Not Like The Answer."

Despite everything the algorithm offered, the district had to scrap the algorithm due to the swift and strong public pushback. As professor Ellen Goodman notes in her 2019 paper "The Challenge of Equitable Algorithmic Change," disgruntled parents carried signs at a school committee meeting that read "families over algorithms," and "students are not widgets."[15]

But the algorithm wasn't really the problem, rather it was the disruptive change to school schedules that was too much, too fast. Implementing the change meant that some elementary school students had bell times that were pulled forward from 9:30 AM to 7:15 AM, some families with children of different ages had to manage several different bus schedules, and some high school students who finished school later had clashes with their extra curricular activities.

Professor Goodman describes the pushback as a case of "algorithmic scapegoating," which Cornell researchers explain is where the algorithm "stood in for substantive issues around equity and disruptive change that were really at stake (though potentially more contentious to discuss) and might well have been at stake even without an algorithm in the picture. The tragedy of the case is that the algorithm could have provided the flexibility to involve the public in choosing among multiple trade-offs. If implemented, it might have created a more equitable system than what existed originally."[16]

The lesson for U.S. regulatory agencies from this episode is twofold: One, algorithmic systems can reduce inequality from human decision-making when they are designed well. Two, communities may not adopt these AI systems even if they could benefit from them if they are implemented in a way that does not explain the rationale behind the use of AI or give citizens sufficient room for recourse.

26. Is the lack of a federal law focused on AI systems a barrier to effective AI accountability?

The barrier to effective AI accountability is not that there isn't a federal law focused on AI but that regulators do not sufficiently recognize or make clear how existing laws in their jurisdiction apply equally to digital and non-digital risks. To this end, it is helpful that regulators across the Biden administration including the Federal Trade Commission (FTC), Civil Rights Division of the U.S. Department of Justice (DOJ), the Consumer Financial Protection Bureau (CFPB), and the U.S. Equal Employment Opportunity Commission (EEOC) recently made clear that existing civil rights laws apply to AI systems and that new laws are not necessary to cover this emerging technology in a joint

[15] Ellen P. Goodman, "The Challenge of Equitable Algorithmic Change.
[16] Ibid.

statement. Other regulators should follow suit by rigorously enforcing existing laws and regulations and conducting a gap analysis to identify any shortcomings.

It is important to understand that AI is a general-purpose technology with many potential applications. Just as a knife is different in the hands of a chef, a soldier, and a surgeon, so too do the risks and benefits of AI depend on how it is being used. Regulators treat knives differently in different sectors, such as creating unique workplace safety standards for scalpels used in hospitals, knives used for food preparation, and knife blades attached to power tools in industrial applications. Likewise, if there is a need for rules, policymakers should create narrow rules for specific AI applications in particular sectors, such as health care and transportation, rather than for AI itself. An AI system to navigate a vehicle should be treated differently than one to automate stock trades or diagnose illnesses, even if they use similar underlying technologies. Forcing all sectors to use the same rules for AI will likely impose excessive or duplicative requirements on some while providing insufficient requirements on others. Creating rules for specific AI applications allows regulators with deeper expertise about particular industries to set appropriate rules for AI applications. For example, insurance regulators may already have considered how to address risks from inscrutable credit scoring models, so whether an insurer uses machine learning models is irrelevant.