Hodan Omaar
Senior Policy Analyst
Information Technology and Innovation Foundation (ITIF)

"The Need for Transparency in Artificial Intelligence"

Before the
U.S. Senate Committee on Commerce, Science, & Transportation
Subcommittee on Consumer Protection, Product Safety and Data Security

September 12, 2023

Chairman Hickenlooper, ranking Member Blackburn, and members of the subcommittee, we appreciate the opportunity to share with you our thoughts on crafting policies to increase transparency in artificial intelligence (AI) technologies for consumers. ITIF is a nonpartisan think tank whose mission is to formulate and promote public policies to advance technological innovation and productivity. In this statement, we offer three considerations policymakers should keep in mind to ensure consumers are protected from harm:

1. While policymakers should encourage companies to adopt the NIST risk management framework, they should recognize that it is not a silver bullet for trustworthy AI. There are a variety of technical and procedural controls companies can employ to mitigate harm and policymakers should encourage companies to explore the full gamut of mechanisms to find those most contextually relevant.
2. Because increasing AI transparency can make some systems less accurate and effective, policymakers should fund research to better understand this tradeoff and evaluate policies for transparency against the impact on system accuracy.
3. Policymakers should hold AI systems to the same standard as human decisions, which are not always transparent.
4. Policymakers should direct NIST to support work on content provenance mechanisms, which are techniques that help users establish the origin and source of content (both AI-generated and human-generated), rather than create policies that simply require systems to disclose when output is AI-generated.

AI offers significant societal and economic benefits in a wide variety of sectors. The biggest risk to consumers is that the myriad opportunities AI offers will not be translated into all the areas where they can make a positive difference in people's lives.

However, there are several other areas of risk to consumers from businesses using AI. One is the creation of unsafe AI products and services, such as a company putting an AI chatbot that advises users to do dangerous things on the market. Another is the use of AI to deceive or manipulate unsuspecting consumers, such as a company using AI to create and spread fake reviews about their goods or services, which ITIF's Center for Data Innovation explores in its

2022 report "How Policymakers Can Thwart the Rise of Fake Reviews."[1] A third is the use of AI to commit crimes that harm consumers, such as using AI to support cyberattacks that steal their sensitive information. While there are other applications of AI that interact with consumers, such as the use of AI to make lending or credit decisions or AI used in employment decisions, we note that these are not in the scope of the subcommittee and therefore keep our comments focused on those that are.

1. While policymakers should encourage companies to adopt the NIST risk management framework, they should recognize that it is not a silver bullet for trustworthy AI. There are a variety of technical and procedural controls companies can employ to mitigate harm and policymakers should encourage companies to explore the full gamut of mechanisms to find those most contextually relevant.

Chairman Hickenlooper and Ranking Member Blackburn are right to state in their recent letter to technology companies that the National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF)—a framework that helps companies identify and mitigate potential risks from AI—can help protect consumers from harm and encourage companies to responsibly develop and use AI.[2] However, it is important to note that many facets of trustworthy AI cannot easily be translated into objective, concrete metrics and technical standards alone are not a silver bullet for trustworthy AI.

For instance, ensuring AI systems are robust and secure is one important element of creating trustworthy AI that protects consumers, and yes, 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).

However, other facets of trustworthy AI, such as ensuring these systems are fair or unbiased, are subjective and cannot be reduced to fixed functions.[3] To see why, consider two e-commerce platforms that use AI algorithms to recommend products to their users. One platform employs an AI system with an objective function to recommend products solely based on customer preferences and purchase history, aiming to provide personalized recommendations without taking into account the price of the products. The other platform uses an AI system with an objective function that considers both customer preferences and product prices, trying to

---

[1] Morgan Stevens and Daniel Castro, "How Policymakers Can Thwart the Rise of Fake Reviews," (Center for Data Innovation, September 2022), https://datainnovation.org/2022/09/how-policymakers-can-thwart-the-rise-of-fake-reviews/.

[2] "Hickenlooper, Blackburn Call on Tech Companies to Lead Responsible AI Use," press release, Apr 19, 2023, https://www.hickenlooper.senate.gov/press_releases/hickenlooper-blackburn-call-on-tech-companies-to-lead-responsible-ai-use/.

[3] 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.

recommend products that not only match user preferences but also fall within the user's budget. Assume both AI systems are designed to be error-free. Even if both AI systems are functioning perfectly, they may have different suggestions for consumers from different socioeconomic backgrounds. Defining which system is more "fair" in this context can be complex, as fairness might involve considerations of affordability, accessibility, and equal opportunity to access desirable products.

This example demonstrates that achieving fairness in consumer product recommendations can be multifaceted and context-specific. Fairness may not have a one-size-fits-all definition. Rather than pursuing technical standards alone, policymakers should be pursuing the principle of algorithmic accountability. As the Center for Data Innovation explains in its 2018 report “How Policymakers Can Foster Algorithmic Accountability,” this principle states 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.[4] 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.

2. Because increasing AI transparency can make some systems less accurate, policymakers should fund research to better understand this tradeoff and evaluate policies for transparency against the impact on system accuracy.

One of the core tenets of transparent AI people cite is explainability. 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. Explanations 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.[5]

---

[4] 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.
[5] “AI Foundational Research – Explainability”, NIST, https://www.nist.gov/topics/artificialintelligence/ai-foundational-research-explainability.

However, it is well-documented that there is often a trade-off between explainability and accuracy. As a 2020 paper led by NIST researcher explains "typically, there is a tradeoff between AI/ML accuracy and explainability: the most accurate methods, such as convolutional neural nets (CNNs), provide no explanations, while more understandable methods, such as rule-based systems, tend to be less accurate."[6]

Policymakers should seek to understand the extent to which this is true for applications that impact consumers and how they can implement policies for increased transparency in a way that does not harm system accuracy. A 2022 paper called "The Non-linear Nature of the Cost of Comprehensibility," published in the *Journal of Big Data* notes that "while there has been a lot of talk about this trade-off, there is no systematic study that assesses to what extent it exists, how often it occurs, and for what types of datasets."[7] It could be the case that high-risk consumer-facing AI applications are more likely to become less accurate if they were made to be more explainable, or it might not. More research would help answer this question. Furthermore, if policymakers want to increase transparency for certain high-risk scenarios, they should also fund research into methods that might limit the impact on system accuracy.

3. Policymakers should hold AI systems to the same standard as human decisions, which are not always transparent.

Policymakers should be careful of holding AI systems to a higher standard than they do for humans or other technologies and products on the market. This is a mistake the European Commission is making with its AI Act. The EU's original proposal contains impractical requirements such as "error-free" data sets and impossible interpretability requirements that human minds are not held to when making analogous decisions.[8] Policymakers should recognize that no technology is risk-free; the risk for AI systems should be comparable to what the government allows for other products on the market.

More broadly, targeting only high-risk decision-making with AI, rather than all high-risk decision-making, is counterproductive. If a certain decision carries a high risk of harming consumers it should make no difference whether an algorithm or a person makes that decision. For example, if it is harmful to deceive consumers by creating fake reviews, enforcement action should be proportional, regardless of whether a human or an AI system was used to create them. To hold algorithmic decisions to a higher standard than human decisions implies that automated decisions are inherently less trustworthy or more dangerous than human ones, which is not the

---

[6] D. Richard Kuhn et al, "Combinatorial Methods for Explainable AI," October 2020, *Preprint: 9th International Workshop on Combinatorial Testing (IWCT 20)*, https://csrc.nist.gov/CSRC/media/Projects/automated-combinatorial-testing-for-software/documents/xai-iwct-short-preprint.pdf

[7] Sofie Goethals et al, "The Non-linear Nature of the Cost of Comprehensibility," March 7, 2022, *Journal of Big Data,* https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00579-2.

[8] Patrick Grady and Kir Nuthi, "The EU Should Learn From How the UK Regulates AI to Stay Competitive," (Center for Data Innovation, April 2023), https://datainnovation.org/2023/04/the-eu-should-learn-from-how-the-uk-regulates-ai-to-stay-competitive/.

case. This would only serve to stigmatize and discourage AI use, which would reduce its beneficial social and economic impact.

4. Policymakers should direct NIST to support work on content provenance mechanisms, which are techniques that help users establish the origin and source of content (both AI-generated and human-generated), rather than create policies that simply require systems to disclose when output is AI-generated.

Some policymakers advocate for policies mandating that generative AI systems, such as those used in customer service, social media, or educational tools, must include notices in their output, informing users that they are interacting with an AI system rather than a human. However, mandatory disclosure requirements may not always be practical or desirable. Many AI applications aim to replicate human capabilities, whether by crafting human-like emails or simulating lifelike customer service interactions. In such cases, labeling content as AI-generated could undermine the very purpose for which consumers use these systems.

Instead, policymakers should support content provenance mechanisms. Content provenance mechanisms are techniques used to trace and establish the origin or source of digital content, whether it's text, images, videos, or any other form of data. For example, one technique is to embed secure metadata within digital files to provide information about the author, creation date, location, and other relevant details. Metadata helps users trace the origin of the content they interact with, whether it's AI-generated, human-made, or a hybrid of both. This approach provides transparency without mandating a disclosure that might compromise the utility of AI systems. It also addresses concerns about the proliferation of misinformation on social networks by allowing users to verify the source of the content they encounter.

The private sector is already conducting work in developing tools and research for content provenance, such as the industry-led Coalition for Content Provenance and Authenticity (C2PA), an initiative that is developing standards and technologies for verifying the authenticity and provenance of digital media content to combat misinformation, establish trust, and increase transparency. The subcommittee should encourage and fund NIST to support and bolster such work.

In conclusion, we appreciate the opportunity to provide our insights on enhancing AI transparency for consumers. Transparency can play a valuable role in achieving algorithmic accountability for some applications and we encourage the subcommittee to support research into when and how this mechanism can be used to support greater use of AI for consumers.