Statement on AI Transparency for the House AI Task Force

Sorelle Friedler, Shibulal Family Professor of Computer Science, Haverford College

Thanks for inviting me to speak about AI transparency today.

I'm the Shibulal Family Professor of Computer Science at Haverford College and the former Assistant Director for Data and Democracy at the White House Office of Science and Technology Policy, where I co-authored the AI Bill of Rights.[1] I have done research on Responsible AI techniques for more than a decade and am a co-founder of the ACM Conference on Fairness, Accountability, and Transparency,[2] one of the main publication venues in the area of transparency and explainability for AI. Before becoming a professor, I was a software engineer at Google.

My research has included techniques to understand how AI systems make predictions. A decade ago, I worked with materials scientists to use AI to accelerate the discovery of new materials. When we were working together, those scientists asked me how it was that the AI system was able to outperform human experimenters when it came to creating new potential materials: what knowledge was the AI system using to make its decisions and how could we learn from that?

These were reasonable questions then and reasonable questions now. We should be able to not only *replicate and transparently describe* each step in a process that uses an AI system, but *explain* how and why it works — this makes us better users of these tools and more effective in getting things done. We developed an AI system that could generate explanations and my colleagues were able to confirm hypotheses derived from it in the lab.[3]

In the decade since that work, the field has come a long way in understanding what works — and what doesn't — for AI transparency and explainability. Beyond its importance to scientific understanding, it has become clear that transparency and explainability — that is, being able to understand *what* AI does and *how* it works - are key consumer needs. This is because AI has an impact on people through:

- automated hiring screening tools used for employment opportunities,
- automated fraud detection tools used in the context of government benefits,
- clinical health care decision support tools used to help doctors better treat patients,
- chatbots and content generators that we interact with daily,

and many other automated or AI-assisted domains, from housing to criminal justice to employment. In cases where important decisions are being made about people, we need to make sure consumers, both individuals and businesses, have transparency into how and why those decisions are made.

---

[1] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[2] https://facctconference.org/
[3] http://dx.doi.org/10.1038/nature17439

AI transparency needs to address three goals:

- **1) safety and efficacy** — these systems should be tested and shown to work before they're used, and consumers should have access to evaluations so they can have confidence those tests were performed;
- **2) fairness** — Americans should be treated fairly by AI, our civil rights and liberties should not be violated by the use of AI systems; and,
- **3) data privacy** — consumers should know what data about them is being used to make an AI-driven decision and be able to correct any errors present in this data.

Without these practices, it can be exceptionally hard to make sense of the information we get from AI systems, and to know whether and how they work.

Take a real-life example (reported by Wired[4]) about one of the most urgent issues facing people in communities across the United States: the opioid crisis.

> A woman in serious pain was in the hospital. Her doctor prescribed opioids, yet after a few days, she was cut off from pain medication, discharged from the hospital, and her doctor terminated their relationship in a letter referencing a "report from the NarxCare database," but no further explanation was given. NarxCare is a database and AI system that is supposed to flag patients at high risk of an opioid addiction or overdose, but she couldn't figure out why it was flagging her, nor could her doctors. It turns out, her dogs' medications had been entered into the database under her name!

This story demonstrates why it is important to know *how* and *why* an AI system makes decisions. First, concerns have been raised by health policy experts about the effectiveness of NarxCare and other opioid overdose risk algorithms[5] that are in wide use. This is a question we should know the answer to, but investigating such systems without direct access to the AI systems or associated data themselves is quite hard. Data on AI system effectiveness should be part of transparency disclosures. Second, if an explanation had been available to the patient or her doctor it might have been clear immediately that the opioid prescriptions raising her risk score were for her dogs — not her — potentially preventing her troubles in the first place!

As this story illustrates, transparency makes systems better and easier for consumers to use. It can also be implemented easily. There are a few different types of transparency mechanisms that are important to consider depending on context.

- **public transparency disclosures**, the type of thing that might be available from a company's About page, can be used to provide detailed information about how a system works, details about evaluations performed that demonstrate efficacy, and contact information for the relevant group at the company that maintains the AI system;
- **notice to individuals** can help inform them that an AI system is in use that impacts them or that they're interacting with an AI system or the results of an AI system;

[4] https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10531142/

- and individualized **explanations** as to how and why an AI system made a decision about a specific individual that can be provided directly to them.

While it's useful for public transparency disclosures to be made for *all* AI systems, individualized notice and explanations are most important for high impact systems that make or assist in decisions about consumers. Consumers impacted by these consequential systems should receive notice that a decision will be made by an AI system *prior to* receiving a decision. And consumers who receive a *negative* decision should receive an explanation. I think of these as akin to **adverse action notices** used in a financial context[6]; explanatory notices that describe the principal factors contributing to an adverse action taken against someone.[7] These would be sent to individuals who are likely already receiving communication about a negative action:

- For example, someone who applies for a job should know at application time if their resume will be screened by an AI-driven system, and if they aren't hired should receive information about the specific reasons the AI system used to screen them out.

Requirements for such notice and explanations could be limited to specific types of consequential decisions, such as financial, employment, housing, or benefits decisions.

When considering what information to include in transparency disclosures it's useful to look to models that have become de facto industry standards,[8] these include:

- **model cards**[9] — developed by Google, model cards include a standard list of questions to be asked about AI models and have been widely adopted by industry as a voluntary transparency mechanism. Think of this as the exam results or report card for the system. And
- **datasheets**[10] — developed by Microsoft Research, datasheets include specific questions that can be asked about the data underlying an AI system. These have been compared to nutrition labels.

Taken together, these transparency disclosures can give consumers an understanding of what an AI system's intended use is — what it should and shouldn't be used for — what factors are used by the model to make determinations, how it was evaluated, what the underlying data is, and so on.

---

[6] https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/

[7] Existing highly used software packages that can produce such explanations include: https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability https://scikit-learn.org/stable/modules/permutation_importance.html https://cloud.google.com/vertex-ai/docs/tabular-data/forecasting-explanations https://github.com/marcotcr/lime https://github.com/shap/shap

[8] See, e.g.: https://modelcards.withgoogle.com/ https://huggingface.co/docs/hub/en/model-cards https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html https://resources.docs.salesforce.com/latest/latest/en-us/sfdc/pdf/salesforce_ai_model_cards.pdf https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/

[9] https://dl.acm.org/doi/10.1145/3287560.3287596

[10] https://dl.acm.org/doi/10.1145/3458723

There are a number of different models for how such AI transparency disclosures can be made available to the public or regulators. Such information could be directly disclosed on a public website, it could be given confidentially to regulators, or it could be written down but kept privately by a company unless formally requested by a regulator. It may make sense to use a combination of these approaches depending on concerns about proprietary information.

Let me be clear, this is information that every good software engineer should have at hand. And it need not be any more a risk to trade secrets than Coca Cola's nutrition label is to their secret recipe.

In closing, there is a lot of possibility here for common sense agreement on basic transparency about AI systems. Individual and business consumers considering paying for systems should have access to information that indicates what an AI system is meant to be used for and whether it actually works.

Thanks.