October 15, 2020

Elham Tabassi,
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

Dear Ms. Tabassi,

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit comments in response to the National Institute of Standards and Technology's (NIST's) request for comment on its draft white paper, "Four Principles of Explainable Artificial Intelligence (NISTIR 8312)," which seeks to develop principles encompassing the core concepts of explainable AI.[1]

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C., 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 important data-related technology trends. The Center is a non-profit, non-partisan research institute affiliated with the Information Technology and Innovation Foundation.

SUMMARY OF COMMENTS

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.[2]

In certain scenarios, some users may also be more likely to trust explainable AI systems. However, there is often a trade-off between explainability and accuracy. In addition, other factors will likely impact trust as well. Indeed, the accuracy and reliability of an AI system is likely to be more important to user trust.

[1] "AI Foundational Research – Explainability", NIST, August 17, 2020, https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability.
[2] Jiawei Su, et al, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, Vol. 23, Issue.5 , pp. 828-841, https://arxiv.org/abs/1710.08864.

Consider two AI systems that predict whether it will rain today. One system is accurate 9 times out of 10, and provides no explanation for its prediction. Another system is accurate 7 times out of 10, and explains which factors (e.g. air temperature, air pressure, wind speed, etc.) it primarily uses to make its assessment. Even though the latter system provides an explanation, users might be less likely to trust it if it is wrong more often.

Moreover, trust is useful, but it is not the only factor that influences adoption. Consumers generally care more about price and quality when making purchasing decisions.[3]

NIST should amend its white paper to clarify the multiple factors that affect trust, particularly accuracy. Moreover, NIST should note the relative dearth of empirical data quantifying the degree to which explainability impacts user trust and user adoption and acceptance of AI technologies.

Finally, since developers do not have the context-specific knowledge to know what will cause harm in a given domain application, NIST should revise their suggestion that systems should be responsible for assessing when they are likely to cause harm.

We offer specific recommended line edits to the draft white paper in the document attached to these comments.

SYSTEM ACCURACY IS MORE IMPORTANT THAN EXPLAINABILITY ACCURACY FOR USER TRUST

NIST's draft white paper paints an overly simplistic picture of the distinction between explanation accuracy (the probability an explanation is true) and decision accuracy (whether a system's judgment is correct or incorrect) that does not capture the various ways these concepts can impact user trust.[4]

For example, 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.[5] 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 on trust was more complex. In highly accurate systems, for example, any explanation, whether the explanation

[3] Alan McQuinn and Daniel Castro, "Why Stronger Privacy Regulations Do Not Spur Increased Internet Use" (Information Technology and Innovation Foundation, July 2018), http://www2.itif.org/2018-trust-privacy.pdf.
[4] Line 211, NIST, "Four Principles of Explainable Artificial Intelligence (NISTIR 8312)" (August 2020), https://doi.org/10.6028/NIST.IR.8312-draft.
[5] Andrea Papenmeier et al, "How model accuracy and explanation fidelity influence user trust in AI" (July 2019), https://arxiv.org/pdf/1907.12652.pdf.

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 performance, a highly accurate explanation had no impact on user trust and a less accurate explanation decreased trust. This example illustrates that at least in some cases, system accuracy is a more decisive factor in creating trustworthy AI than explanation accuracy is. NIST already highlights resiliency, reliability, bias, explainability, and accountability as properties that characterize trust in AI systems, but it should add decision accuracy to this list, and be clear that while explanation accuracy can affect user trust, it is not necessarily as important as other factors, such as system accuracy and reliability.

More importantly, the 2019 study showed that users did not trust an inaccurate classifier, regardless of the accuracy of the explanation given. This finding suggests that attempts to mislead users through inaccurate explanations, as discussed in the draft white paper, may be difficult for highly accurate systems.

CONSUMERS CARE MORE ABOUT PRICE AND QUALITY THAN ETHICAL DESIGN

NIST takes at face value the assumption that if AI systems are not explainable, they may cause users to be suspicious that the system is biased or unfair which "may slow societal acceptance and adoption of the technology, as members of the general public oftentimes place the burden of meeting societal goals on manufacturers and programmers themselves."[6] But this presupposes that when making purchasing decisions, consumers care more about whether a system is biased or unfair than they do about its price or quality. Yet there is virtually no evidence suggesting this to be the case.[7]

For example, a survey from the Center for Data Innovation found that only 19 percent of Americans agreed with the statement, "If I am buying a smart toaster (i.e. a toaster controllable by a mobile app), I am willing to pay more for one that is certified as 'ethical by design.'"[8] This shows that while some consumers may pay lip service to ethical design, this does not match their behavior which is a

[6] Line 128, NIST, "Four Principles of Explainable Artificial Intelligence (NISTIR 8312)" (August 2020), https://doi.org/10.6028/NIST.IR.8312-draft.
[7] Daniel Castro, "Europe will be left behind if it focuses on ethics and not keeping pace in AI development," Euronews, August 7, 2019, https://www.euronews.com/2019/08/07/europe-will-be-left-behind-if-it-focuses-on-ethics-and-not-keeping-pace-in-ai-development.
[8] Daniel Castro, "Bad News, Europe: Consumers Do Not Want to Buy an "Ethical" Smart Toaster" (Center for Data Innovation, March 2017), https://www.datainnovation.org/2019/03/bad-news-europe-consumers-do-not-want-to-buy-an-ethical-smart-toaster.

more objective measure of trust. Similarly, few consumers, other than those who perhaps took auto repair classes in high school, know how their automobile works. They simply trust that their vehicle's complex systems, such as the electronic ignition, fuel injectors, and anti-lock brakes, will work as expected.

NIST should clarify that in terms of societal acceptance and adoption, explainability and its impact on trust is not necessarily as important as other attributes of an AI system, such as how much it costs or how well it performs, and the need for more research on this relationship.

SYSTEMS SHOULD NOT BE RESPONSIBLE FOR ASSESSING WHEN THEY CAUSE HARM

NIST's proposal says that AI systems should explain when they have reached their knowledge limits, meaning AI systems should "identify cases they were not designed or approved to operate [in], or [cases in which] their answers are not reliable." But this requirement incorrectly conflates the responsibilities of system developers, who create AI systems, and system operators, who are responsible for deploying AI systems.[9]

For example, a government agency that uses an algorithm to screen people at border crossings, or a company that deploys an AI system to vet job applicants, are operators, while a developer who publishes an algorithm that classifies different datasets is not. This is important because simply creating an algorithm that can be applied to situations where it exhibits some kind of demographic bias does not cause harm in itself and should be of no concern unless an operator applies it in a way that could cause harm.[10]

By suggesting systems be responsible for assessing when they are likely to cause harm, NIST wrongly assumes developers can predict or control for every possible harmful outcome that could arise from the use of their algorithms. In reality, this is near impossible. Developers do not have the context-specific knowledge to know what will cause harm in a given domain application. For example, what constitutes harm in consumer finance involves dramatically different criteria than what constitutes harm in healthcare. Only an operator can verify a system acts "under [the] conditions for which it was designed" or identify when "the system reaches a sufficient confidence."[11] NIST should differentiate

[9] Line 230 - 231, NIST, "Four Principles of Explainable Artificial Intelligence (NISTIR 8312)" (August 2020), https://doi.org/10.6028/NIST.IR.8312-draft.
[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.
[11] Line 169 - 170, NIST, "Four Principles of Explainable Artificial Intelligence (NISTIR 8312)" (August 2020), https://doi.org/10.6028/NIST.IR.8312-draft.

between these responsibilities and focus solely on explainability, rather than accountability, in this white paper.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
dcastro@datainnovation.org

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