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

All comments will be made public as-is, with no edits or redactions. Please be careful to not include confidential business or personal information, otherwise sensitive or protected information, or any information you do not wish to be posted.

Comment Template for First Public Draft of Four Principles of Explainable Artificial Intelligence (Draft NISTIR 8312)

Submit comments by October 15, 2020 to:
explainable-AI@nist.gov

Comment #	Commenter organization	Commenter name	Paper Line # (if applicable)	Paper Section (if applicable)	Comment (Include rationale for comment)	Suggested change
1	Center for Data Innovation	Hodan Omaar Daniel Castro	124	Introduction	The footnote associated with this sentence references the Fair Credit Reporting Act (FCRA) which regulates the collection of consumers' credit information and access to their credit reports. This indicates that the FCRA requires consumer reporting agencies to share the rationale behind their decisions. However this is not the case; the FCRA does not require consumer reporting agencies to share the rationale behind their decisions.	Remove the reference to the Fair Credit Reporting Act from footnote 8.
2	Center for Data Innovation	Hodan Omaar Daniel Castro	125 - 126	Introduction	This sentence states that a lack of explainability can negatively affect the level of trust users will grant an AI system. While this is technically true, this sentence does not reflect the reality that in many cases a lack of explainability can increase trust, especially in highly accurate systems, as explained further in our comment #7.	Revise this sentence to qualify that the following statement is only true in some cases: "the failure to articulate the rationale for an answer can affect the level of trust users". Include references to literature, such as Papenmeier et. al [2019], which evidence that the relationship between explanability and user trust varies across accuracy levels.
3	Center for Data Innovation	Hodan Omaar Daniel Castro	128 - 132	Introduction	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." 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. 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.'" This shows that while some consumers may pay lip service to ethical design, this does not match their behavior which is a more objective measure of trust.	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.
4	Center for Data Innovation	Hodan Omaar Daniel Castro	134	Introduction	This sentence highlights resiliency, reliability, bias, and accountability as the properties, besides explainability, that characterize trust in AI systems. It does not include decision accuracy which is a more important factor than explanation accuracy in increasing user trust as per our comment #7.	NIST should include decision accuracy to the list of properties that characterize trust.
5	Center for Data Innovation	Hodan Omaar Daniel Castro	166	Four Principles of Explainable AI	This sentence defines meaningful AI as a function of individual users and their prior knowledge, implying that if two individuals were to fall within the same broader group, e.g. doctors, the system will be more meaningful for the doctor who has greater prior knowledge. This does not align with the explanation of meaningful AI given in section 2.2 which says: "Multiple groups of users for a system may require different explanations. The Meaningful principle allows for explanations which are tailored to each of the user groups." The discrepancy between whether the meaningful principle is intended to enable explanations for individuals or user groups creates confusion.	NIST should clarify how granular explanations need to be in order to fulfil the meaningful principle, meaning it should define whether explanations need to be understood at the user group level or the individual level. However, greater explainability often imposes, at a technical level, limits on system complexity and system performance. NIST should caution against describing meaningfulness as explanations for individuals as this may have impacts on system performance which is a more decisive factor in creating trustworthy AI, as explained in comment #7.

6	Center for Data Innovation	Hodan Omaar Daniel Castro	169 - 170	Four Principles of Explainable AI	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. 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. 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."	NIST should differentiate between developer and operator responsibilities and focus solely on explainability, rather than accountability, in this white paper.
7	Center for Data Innovation	Hodan Omaar Daniel Castro	211 - 214	Explanation Accuracy	This section 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. 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. 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 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.
8	Center for Data Innovation	Hodan Omaar Daniel Castro	224 - 225	Knowledge Limits	This sentence states that a system may be considered explainable if it can generate more than one type of of explanation. This broad definition does not refer to properties of trustworthy systems noted in line 134 of the draft, including resiliency and reliability. It also does not refer to system accuracy which is an important element of trustworthy systems as we have explained in comment #7.	NIST should update the definition of what is considered an explainable system and qualify it in terms of accuracy, reliability, and resilience.
9	Center for Data Innovation	Hodan Omaar Daniel Castro	233 - 234	Knowledge Limits	This sentence states that one purpose of the knowledge limits principle is to increase trust in a system by preventing misleading, dangerous, or unjust decisions or outputs. This does not align with the purpose described in line 143 of this draft which states principles are given to provide a baseline comparison for progress in explainable AI. This sentence conflates accountability and explainability.	NIST should redefine this principle, focusing solely on explainability, rather than accountability.

10	Center for Data Innovation	Hodan Omaar Daniel Castro	245	Types of Explanations	This section intends to describe five types of explanation, but instead, describes five circumstances under which an explanation may be given: to inform a user; to generate trust and acceptance; to assist with audits for compliance and regulations; to facilitate developing, improving, debugging, and maintaining of an AI algorithm or system; or to benefit the operator of a system. While this information is useful, the title is misleading. Further, an explanation of different types of explanations is missing in this document.	NIST should change the title of section 3 to clarify it describes the circumstances under which an explanation may be given. It should also include a new section that describes the types of explanation that an AI system may provide to a query. Aristotle's Four Causes model, also known as the Modes of Explanation model, may serve as a foundation for this section. It states four types of 'causes' (that translate today as 'explanation') that can be used to provide answers to 'why' questions: 1. The material cause of a change or movement: The substance or material of which something is made. For example, rubber is a material cause for a car tire. 2. The formal cause of a change or movement: The form or properties of something that make it what it is. For example, being round is a formal cause of a car tire. These are sometimes referred to as categorical explanations. 3. The efficient cause of a change or movement: The proximal mechanisms of the cause something to change. For example, a tire manufacturer is an efficient cause for a car tire. These are sometimes referred to as mechanistic explanations. 4. The final cause of a change or movement: The end or goal of something. Moving a vehicle is an efficient cause of a car tire. These are sometimes referred to as functional or teleological explanations. As Tim Miller from the University of Melbourne describes in his 2018 paper
11	Center for Data Innovation	Hodan Omaar Daniel Castro	327 - 333	Overview of Principles in the Literature	This section explores a paper from Wachter et al. that claims counterfactual explanations are sufficient. The key insight from this paper and from others is that people do not explain the causes for an event per se, but explain the cause of an event relative to some other event that did not occur; that is, an explanation is always of the form "Why X rather than Y?" This finding is significant as it may imply AI systems need only provide counterfactual explanations. There is a great amount of research in the philosophical and cognitive science literature that supports this claim. NIST should include more of this research in this section that provides an overview of the literature.	NIST should include more research on counterfactual explanations such as: - P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27 (1990) - J. Van Bouwel, E. Weber, Remote causes, bad explanations?, Journal for the Theory of Social Behaviour 32 (4) (2002) - G. Hesslow, The problem of causal selection, Contemporary science and natural explanation: Commonsense conceptions of causality (1988) - D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin 107 (1990)
12	Center for Data Innovation	Hodan Omaar Daniel Castro	417 - 418	Self-Explainable Models	This sentence states that "many sources discuss an accuracy-interpretability trade-off," yet the draft paper does not include sufficient discussion of this trade-off or include what these sources have found. The trade-off between accuracy and interpretability has great implications, as discussed in our other comments, so it is important that NIST states this trade-off clearly and discusses its implications.	NIST should include details of the the findings from the sources it cites in this sentence. Given this section is an overview of the literature in this space, it should include these here.
13	Center for Data Innovation	Hodan Omaar Daniel Castro	524 - 527	Adversarial Attacks on Explainability	This section discusses adversarial attacks on explanations, claiming that explanations "without 100 percent accuracy" are at risk of being attacked. However the 2019 study by Papernmeier et al. showed that users did not trust a bad classifier, no matter the explanation given. This illustrates that system accuracy is important for trust. For highly accurate systems, adversaries may find it difficult to mislead users through inaccurate explanations.	NIST should include the importance of accuracy in addressing threats of adversarial attacks in this section.