IEEE SA Comments for the 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	IEEE	IEEE-USA AI Policy Committee		N/A	General Comment: The document as constituted raises concerns that may limit its uptake among policymakers and designers alike.	Please see comments below.
2	IEEE	IEEE-USA AI Policy Committee		Section 2	We expected the document to explain how the principles could be used to measure and evaluate explanations generated by AI systems.	We recommend Section 2 focus on how these principles can be used to measure and evaluate explanations in practice.
3	IEEE	IEEE-USA AI Policy Committee		Section 2	The four principles of explainability are at times overlapping, inconsistent, and ill-defined.	We recommend replacing abstract principles with what design elements are necessary to ensure that a useful explanation is available (similar to the "Types of Explanations" recommendation below): Inputs: the system should be able to explain what data the AI consumed or synthesized, the sources and provenance of the data, the assumptions and limitations* of that data and its relevance and reliability as applied to the current task. Processing: the system should be able to explain what methods are used to transform the inputs into outputs, the assumptions and limitations* of those methods and their relevance and reliability as applied to the current task. Outputs: the system should be able to explain the accuracy, significance, confidence, uncertainty in the output, and the assumptions and limitations* of those methods and their relevance and reliability as applied to the current task. Where relevant, the system should be able to explain which inputs and processes were relied upon; which aspects of the inputs and processing were authorized, verified and validated, by something other than the AI system itself; and any error messages. Interaction: the system's explanations should be delivered accounting for timing, context, level of detail and abstraction, criticality of the decision being made or informed, communication methods such as visualizations, hypothesis testing, counter-examples, and whether the system is operating within its bounds. *Assumptions and limitations may require further description in each of the inputs, processing, and outputs. In
4	IEEE	IEEE-USA AI Policy Committee		Section 2	The four proposed principles do not explain why we really need explainable systems. We suggest that the document explain that the goal of explainability is to calibrate people's trust appropriately. In other words, the real goal of explainable AI is not that it is accurate, or evidence-based, or meaningful, it is that the system helps achieve the stakeholder's goals.	We recommend the document should include the following: "The principles of explainability are not an end in themselves. The principles are the means to designing explainable AI systems that are worthy of the stakeholder's trust. In designing AI systems for deployment, there will likely need to be tradeoffs in how to achieve the elements of explainability above. Therefore, designers must understand the stakeholders goals, which include understanding: (1) the technical processes of the AI system; (2) the AI system's decision-making process, particularly when an AI system's decision has a significant impact on people's lives; (3) the fit of the system to the current task, including necessary and appropriate notifications, directed to the users, who review such automated decisions in order to avoid complacent findings and accepting results without any further considerations; and, (4) the information necessary to enforce a legal right or privilege, such as a right to appeal, where it exists."

5	IEEE	IEEE SA	208	Section 2.3	"Explanation Accuracy" seems to impose accuracy on a system's explanations, but does not fully take into account the possible real-world outcomes. A toy could accurately explain its actions, such as: "I am going to play with you now," but designers' explanations of how something is generated, without context, can miss the mark and leave room for miscommunication and potential harm. Sociologists and anthropologists could say, "why does the toy have a female voice versus male? Does this denote only girls can play or that the toy is female and therefore subservient?" Without including a full definition of what the term "accuracy" means leaves much room for miscommunication and potential harm.	Recommend: Fully define the term "accuracy" and what needs to be included in an "accurate" explanation.
6	IEEE	IEEE-USA AI Policy Committee		Section 3	Section 3 on the "Types of Explanations" should be replaced because the statement should identify any required "Types of Explanations" and describe them in terms that do not presume who needs what type of information. The current framing is designed around each stakeholder instead of the system itself. To aid design, it is easier to say what the system ought to be able to provide through design then allow strategy and compliance to characterize who receives the information. Additionally the current definitions of the types of explanations are circular use cases - user benefit (informs the user); societal acceptance (generates trust and acceptance by society); regulatory and compliance (assists with regulations and safety standards); system development (assists maintaining a system); and owner benefit (benefits the owner). They are both overlapping and inexhaustive.	We recommend striking and replacing "Types of Explanations" (Section 3, Lines 247-276 - starting with "To Illustrate") with the following: "There are at least five types of explanations that each system must be able to provide relative to the system specifications -- though different levels of detail may be provided to different stakeholders such as users, regulators, system developers, or owners and operators: (1) Rationale explanation: This type of explanation describes the reasons that led to a decision, delivered in an accessible and non-technical way. (2) Responsibility explanation: This type of explanation describes who is involved in the development, implementation, management and operation of an AI system, (this is particularly important for operators who often are non-experts in AI) and who to contact for a human review of a decision. (3) Data explanation: This type of explanation describes what data has been used in a particular decision and how; what data has been used to train and test the AI system and how (if the data are unbiased and fair) the quantity and quality of data is sufficient. (4) Safety explanation: This type of explanation describes the evidence that the system is accurate, reliable, secure and robust. (5) Impact explanation: This type of explanation describes the impact that the use of an AI system and its decisions has or may have on an individual, and on wider society."
7	IEEE	IEEE SA	Line 289	Section 3	This section is helpful. We suggest a fail-safe alternative to any system where a group of users may need an explanation beyond what a system was designed for. Typically this is a human customer service representative. Note the logic is to also have the human at the end of these processes ask users which portion of the instructions they received they did not understand. These insights in aggregate can be used to inform / update the level of explanainbility for future use to improve clarity.	
8	IEEE	IEEE SA	Line 313	Section 3	The document lacks mention of user "agency." Such agency is a key to all apsects of disclosure, explainability, etc. Designers cannot assume that even with best intentions or all of the factors listed here for explainability, etc., that a user not given agency will ever understand how a product will actually affect them in use in the context of their lives. For instance, how would someone using a voice assistant for the first time fully understand what "explainable" means without living with the device for a week or more? While this may make this text seem even more complex, it is a form of user or customer testing which is often ignored regarding algorithmic systems, in lieu of pushing products to market quickly.	

9	IEEE	IEEE-USA AI Policy Committee		Section 6	With respect to Section 6, if retained, we suggest that the section include an exhaustive literature search resulting in a comprehensive bibliography but with only the main results summarized in the body of the document and indications of where there is consensus and where consensus is lacking. However, since it is not clear that even such an improved version of Section 6 is necessary, our alternative recommendation is to remove it entirely. If the purpose of the document is to provide objective and actionable explainability standards that can be used to measure whether particular explanations generated by AI systems are adequate, then it should be possible to define such standards on their own terms, without reference to the psychological limits and barriers that humans face when attempting to explain their own decisions. Furthermore, even if it is desired to use human strengths and weaknesses at explaining their own decisions as a basis of comparison to explanations generated by AI systems, the current version of Section 6 draws on a very limited and non-representative set of literature. For example, Section 6 does not acknowledge or cite the extensive evidence in legal practice or human factors or expert decision making research that demonstrates the ability of humans to provide accurate explanations of their own decisions in certain contexts and when adhering to certain practices. Instead, Section 6 refers almost entirely to evidence supporting the conclusion that humans are unable to provide accurate explanations of their decisions. This is at odds with a more balanced and appropriate review of human psychology. A version of Section 6 that could provide a useful description of human explanatory skill as a basis for comparison to AI explanatory skill would require a more extensive and balanced review of the relevant literature. As it stands now it is not clear why the authors selected the	We recommend rewriting, and in the alternative, striking Section 6.
10	IEEE	IEEE SA	Line 584	Section 6	While the explanation of human and AI systems is helpful in one regard, it avoids the seminal aspects of algorithmic bias and fundamental systemic-level issues such as how systemic societal racism or other bias influence both a human and any AI system that human may use. Here, things like human intution, emotional intelligence, and experience in a profession also come into play. Any AI system needs to be examined not only with a question of "accuracy" in terms of how data is delivered from a computational standpoint, but from a socio-technical standpoint by those professions or vertical experts beyond those who created the systems themselves (e.g., sociologists, anthropologists, etc.)	
11	IEEE	IEEE SA	Line 699	Section 7	Problems could arise from the introduction of principles where humans have a limited capacity to reach them as compared to AI / machines. If one human cannot provide an explanation of a system to another human, this is concerning. While it is a given that specific outputs of a calculation or how an algorithm was formed may be beyond the scope of a human without expertise around those issues, the focus of Explainable AI by and large should be on the end user. Where end users are people, we need to pick principles where those creating AI favor end users over any design that favors the designers in isolation.	

12	IEEE	IEEE-USA AI Policy Committee			We are concerned that the language throughout the document mistakenly relies on myths regarding how AI systems operate.	We recommend that the document address or offer the following as guidance to designers in the development of AI systems in its own section. Users do not always need to be convinced to trust AI systems. Having a trustworthy system is a predicate not a presumption. There are plenty of AI systems that are not worthy of being trusted. Having users trust untrustworthy systems is a source of failure that should not be encouraged. Recommendation: Designers should think about how to identify and communicate when their system should not be trusted. The process is not entirely automated as the document suggests. Many systems include AI systems working in cooperation with humans, where the AI performs some functions and humans perform other functions. Therefore, the “explanation” of interest may be regarding the inputs, processes, and outputs of the human-machine system, not merely of the AI system alone. Recommendation: Consider expanding the definition of “explanation” to cover inputs, processes, and outputs of the human-machine system rather than merely the AI system alone. In some cases it may be impractical or not useful to provide a low-level explanation for every “decision” that is made (e.g., AI systems that are making high-frequency “decisions”--lane-assist, real-time video processing, etc.). Instead systems may need to provide explanations of the continuous set of AI decisions. Recommendation: Designers should consider what level of abstraction explanations will need to be for each user.
	IEEE	IEEE SA	Overall comment		We recommend the work of Sandra Wachter for all these terms (explainable, etc). Her work with Luciano Floridi is a seminal paper: https://discovery.ucl.ac.uk/id/eprint/10038294/1/Wachter_Transparent_explainable_accountable_AI.pdf	