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	University of Maryland	Aaron M. Roth	265	3. Types of Explanations	The paper discusses how using explanations for Debugging is useful during system development. In a related but separate type of use case, explanations can be used for troubleshooting. For example, Adaptive Cruise Control on a car might fail when the camera views are obscured or blurred by rain or snow. This is not a bug in the system, but impairs operation, and the system informs the user of this.	Add this as an example type of explanation, whether as an additional example under User Benefit or as an additional type.

2	University of Maryland	Aaron M. Roth	160-163	2. Four Principles of Explainable AI	Some examples of outputs are given: decision, ordered recommendation, detections/highlights. Another important type of output is an action. Whereas a decision usually implies some information that a human will take and then, based on that information, engage in action, many AI models output actions directly that an agent then takes. Stock-trading algorithms and many robotics applications such as self-driving vehicles are examples of this.	Consider adding “action” as an example type of output. I am aware that the list of examples is not intended to be exhaustive, but it could be a useful addition for the reader.
3	University of Maryland	Aaron M. Roth		2. Four Principles of Explainable AI	Relevant to real-world applications such as robotics, factory machines, and self-driving vehicles, there is missing a discussion of Pre-Runtime Verifiability and analysis. This is the idea that, given a model intended to be executed and produce decisions as output which will be immediately executed, there is an interest in determining the degree to which it will operate safely (or operate safely given certain assumptions). We want to be able to produce explanations not after the fact, as in the case where there is still a human intermediary between output and real-world action (as in the case of a medical or legal recommendation system), but ahead of time, to catch potentially dangerous eventualities before they occur. This is another category of explanations, which could be described as "describing and categorizing the nature of possible outputs given certain situations ahead of time" which is very important in a number of real-world applications and not explicitly encompassed by the document currently. Performing these kinds of analysis and verification is a type of explainability that does not seem to be covered by the existing four principles (or at least, if so, is not articulated).	Add a fifth principle of “Verifiability” or “Pre-Runtime Verifiability/Analysis” to the document. (Potentially, the topic to be added could also be a sub-part of Known Limits, or as a type of explanation that could be produced, although it seems to be different from both of those.) This principal measures the degree to which a model is going to produce output in a predictable manner, or in a manner subject to certain constraints. Different types of AI algorithms interact with this principle in different ways. Theoretically, a self-explainable model should be able to be fully verified ahead of time (although it could be difficult in practice). Other types of explainable AI algorithms may have more difficulty providing such guarantees.

4	University of Maryland	Aaron M. Roth	709		suggestion	As an additional nod to relevant topics beyond the scope of the paper itself, the document could mention that AI could be used to explain human rationale in situations when a human cannot explain their own action correctly.