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		Aman Patel			The four principles outlined in this paper seem to be missing a major component. A self-driving car AI system might hit a pedestrian in order to avoid crashing into another car. In an audit of such a scenario, the system might produce an explanation similar to the following: "a leftward swerve was made in order to avoid car-to-car collision." This explanation fits all four principles outlined in the paper. It delivers accompanying evidence for its output, which in this case is the steering of the car; that explanation is understandable by individual users; it correctly reflects the system's process for generating that output; and the system was operating under conditions in which it was designed for. These four principles are satisfied, but the explanation is clearly not sufficient for our expectations. We must still know why the system chose to hit the pedestrian instead of the car--was it programmed to place a higher priority on avoiding cars than on avoiding pedestrians? Did it reason that there was a lower chance of hitting the pedestrian than hitting the car? Did it even recognize the pedestrian? On first sight, this looks to be categorized under the "accuracy" principle, but I would argue otherwise. A system may give a complete and fully accurate explanation of all the factors that went into its decision, but it is also important that we are able to determine the factors that were not part of its decision, but should have been. This is especially true for systems which are designed to make decisions in multiple paradigms, with many separate streams of input, knowledge, and output. We might call this principle "comprehensiveness," because it involves gaining a clear picture of what the system's inputs were, what information it had, and what information it did not factor into its decision-making process.	Inclusion of a fifth principle, which specifies the explanation be *comprehensive*

2		Aman Patel			However the explanation is generated, the system should always present enough information to satisfy all interpretations of the query at hand, at every level of abstraction on which it can operate. If the self-driving system was able to recognize humans and cars as separate categories of objects, it should be able to specify that it was avoiding a car, not just an "object in close proximity," or at an even lower level of abstraction, a "region of high values in a proximity heatmap." This is likely subsumed by the "meaningful" principle, but it might be useful to specify that explanations should be meaningful at different levels of abstraction and modeling.	Inclusion of a mention that explanations should be meaningful not only to different audiences, but also on all levels of abstraction at which the system operates.