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	Leidos	Thomas Boggs	160	2	"Output" is identified as a key term and defined as "the result of a query to an AI system." This is an unnecessarily limited definition, which doesn't cover applications such as RL agents, autonomous vehicles, robotic assistants, etc., whose actions are not limited to user queries.	Expand definition to include responses, actions, and decisions independent of queries.
2	Leidos	Thomas Boggs	241	2.4	The example provided is relevant but is not an example of "knowledge limits" per se; rather, it is more like an example of "information limits" (the input does not provide adequate information from which to make an accurate/credible prediction).	This section currently identifies two categories of limitations: out-of-domain and knowledge limits. Expand it it to include the third category of information limits. The blurry bird example is really an example of an information limit. An example of a knowledge limit would be an image of a type of bird that the system was not trained to recognize. In such an example, the input is still in the domain of the system (i.e., birds) and sufficient information is provided (adequate image content) but the system cannot make a reliable decision because the input does not belong to one of the N bird species the system has learned to identify.

3	Leidos	Thomas Boggs			3	The "types" of explanation presented seem to be more like "purposes" of explanation and some of those seem redundant.	Provide a list or taxonomy of actual explanation types (feature weights, natural language explanations, attention maps, etc.). Providing a matrix with types and purposes along the two axes to indicate which types are appropriate for which purposes would be beneficial.
4	Leidos	Thomas Boggs		324	4	This paragraph is rather contentious. Deep networks are used and will continue to be used for "high stakes" decisions. While explainability is desirable and sometimes necessary, it isn't the only path to Trust in AI. Exceptional performance and robust T&E can result it trustable blackbox models.	Remove paragraph or provide more balanced discussion.
5	Leidos	Thomas Boggs		503	5.3	The minimal change to an input required to change an output is not always desirable for explainability. By that criterion, a single-pixel adversarial attack to an image would be considered a counterfactual explanation. But that clearly would not provide the type of information one would desire from an explanation (though it does provide other information wrt resiliance of the model).	Provide discussion of contexts in which different types of explanation are relevant.
6	Leidos	Thomas Boggs			6	This is a great section for setting the context for AI Explainability and its challenges. It implicitly highlights that we often apply double-standards and/or unreasonable expectations wrt AI Explainability (i.e., we expect AI to provide explanation in a way that humans themselves often can't provide).	

7	Leidos	Thomas Boggs		General	"Explanation" as a fundamental Principle of Explainable AI comes across as a rather weak in the way presented, since it really just amounts to a basic requirement (expressed as a binary "yes/no" attribute).	Rather than presenting a binary property of AI systems (it either explains or it doesn't), consider expanding this principle to something like "Completeness", which relates to the level or extent to which decisions are explained.
8	Leidos	Thomas Boggs		General	Numerous concepts related to Explainability (and AI in general) are used but not defined. For example, "Explainability" is not contrasted with "Interpretability".	Either provide a glossary of key AI/Explainability-related terms or highlight their definitions in the document. This would be helpful for establishing a common lexicon of terms in the community.