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.

#	Paper Line # (if applicable)	Paper Section (if applicable)	Comment (Include rationale for comment)	Suggested change
1	General comment	General Comment	Meaningfully explainable AI requires more than just explanations on the algorithms outputs. Often, information about the provenance of training datasets and individual data points, as well as information about collection methods, can be very helpful for understanding the output of a system. For example, if an image recognition system is trained to identify cardinals, but is only given training images of a cardinal that were captured in summer in low resolution and some similar looking species' pictures were captured in winter and higher-resolution, this may help "explain" why that system incorrectly identifies cardinals as that other species when shown in a winter setting. Likewise, knowing if supervised labels in a dataset came from a particular special interest group's perspective on the topic can elucidate systematic understandings of the way the algorithm might replicate that perspective.	We recommend NIST include discussion of the importance of providing transparent information and metadata about how a system was developed and trained. For example, NIST may cite efforts like Google model cards or the ABOUT ML project at the Partnership on AI and their value to improving explainable AI.

2	125-132	1	In addition to the way explainable AI can “influence public perception of the system,” it can also help improve the safety of AI systems. This is especially true if the system adheres to the “Knowledge Limits” principle of explainable AI. By reducing the “black box” effect, AI developers and end-users of systems can more easily identify possible limitations of the algorithm before they cause significant harm.	Recommend adding a new sentence following the sentence ending in “perception of the system.” on line 132 that states: “Further, the explainable outputs of systems may help recipients identify underlying problems in the algorithm or the training data and thus improve the resiliency, reliability, and accountability of the system.” This will allow the paragraph to flow into the next paragraph, and highlight the value of explainable AI to the development of safe AI systems.
3	160-163	2	This definition of “output” should be expanded to better capture the functions of AI agents, in addition to those from analytical recommendation systems. The current definition suggests outputs only come from a “query to an AI system.” However, AI agents do not produce a single output in response to a query, rather, they produce a series of actions in response to provided command or goal. Thus, the “outputs” of AI agents should include the stream of consequential actions and decisions an agent may make in its environment.	Edit line 160 to say (italics represent existing text): *The output is the result of a query to an AI system or the consequential actions* taken by an AI agent in response to a given command or goal. Include a new example following the end of line 163, stating: “For an autonomous vehicle, the output is the series of actions it takes in response to a driver’s command.”
4	229-244	2.4	We strongly support NIST’s insightful inclusion of “Knowledge Limits” as one of the four principles. We believe this principle, as helpfully articulated by NIST, is often neglected in other discussions about explainable AI, and it is perhaps the most important principle offered in the document. Knowledge Limits is essential for the safe and ethical implementation of AI systems into the real world	
5	375-379	5	It should be mentioned that more useful variants show the top few such counterfactual examples for different classes or features.	Following the end of the sentence in line 379, “...different decision” add a sentence that says “Some systems can also produce multiple counterfactual examples for different classes or features, which can provide a more meaningful explanation of the output.”

6	399-402	5	While it is true that there is limited research measuring explanation accuracy, the research highlighted in section 5.4 (Adversarial Attacks on Explainability) can likely be repurposed and extended to develop metrics on how accurate vs. misleading an explanation is to a recipient.	NIST should add a sentence following “...how the trained models differ” on line 404 to say: “Continued research on adversarial attacks on explainability, as discussed later in Section 5.4, can also be repurposed to provide techniques that can be used for quantification of explanation integrity.”

7	403-406	5	We believe this section is critical and can be improved with further discussion. First, NIST should consider directly highlighting the need for additional investment and research “on developing algorithms that understand their knowledge limits” (403-404). Second, NIST can identify some relevant subfields in the AI literature, such as confidence representation, metauncertainty, distributional shift detection, out-of-distribution detection, open world reasoning, and declarative ontologies. Communicating to a user the nature and size of the uncertainties that bore on an output helps them contextualize and calibrate usage of the system overall as well as each of its outputs individually. In the case of a typical learned classifier, though it implicitly respects the open-world assumption regarding properties of the instances it may encounter, it implicitly enforces an unreliable closed-world assumption (https://link.springer.com/chapter/10.1007/978-1-349-13277-5_4) regarding the semantics of its inference and outputs: i.e. a tautological disjunction of its training classes. Many modern learning algorithms can be expressed as	We recommend rewriting and replacing lines 403-406 with the following. “Algorithmic systems that understand their knowledge limits and declare when a validly-formatted data input is out of their scope are not prevalent today, but research from multiple subfields of AI can be brought to bear to meet this principle. Knowledge limits have a significant history of being addressed in the literature, but not often under the term 'knowledge limits', and not as a single subfield. For example, research on quantification of uncertainty (https://ieeexplore.ieee.org/document/8825816), metareasoning about uncertainty (https://www.ijcai.org/Proceedings/15/Papers/229.pdf), quantification of ambiguity (https://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-03/SS08-03-002.pdf), and manifold characterizations (https://arxiv.org/pdf/1805.11783.pdf) are directly relevant to modeling knowledge limits. The framing of knowledge limits as one concept and as a principle for explainable AI promotes a cohesive understanding and communication of a system's limits. Systems that can clearly communicate where their uncertainties are have been shown to help users calibrate and build their trust (https://www.cell.com/patterns/pdf/S2666-3899(20)30060-X.pdf) in such systems rapidly. In the simplest case, it is common for models to output real-valued probabilities or scores rather than hard decisions, which reflect the algorithms’ confidences in their predictions. Just giving such a real-valued output or even a probability, however, should be considered inadequate to meet this principle because many

8	491	5.2	It would be helpful to include mention of and citation to global explainable AI algorithms for deep reinforcement learning systems.	Add discussion and reference to the research on this topic. For example, see new research: Heuillet, A., Couthouis, F., Díaz-Rodríguez, N. (2020). Explainability in Deep Reinforcement Learning, https://arxiv.org/pdf/2008.06693.pdf
9		6	We agree with NIST that it can be useful to consider how humans do and do not provide acceptable explanations for "outputs" in line with the four principles. As NIST notes, this is especially important for developing a "better understanding of the dynamics of human-machine collaboration" (line 563). However, we are concerned that NIST's description of the poor ability of humans to explain decisions may create the false impression in the reader that similar systemic problems would be acceptable in an AI system. Imperfection in human explanation should not excuse poorly designed, unexplainable AI systems. We are particularly concerned by the notion that humans serve as a "benchmark" for explainable AI, where a benchmark may be mistaken by a reader to imply that human-level is an acceptable threshold or standard for explainable AI. Rather, human-level performance should be a "benchmark" but not an acceptable standard unto itself, which is a distinction that may be lost on some without clarification. Further, NIST should indicate that alternate benchmarks to human-level performance are possible and perhaps most important to develop and use.	Include a new sentence following " …conclusions are largely unreliable" (line 561) that reads "Though it is helpful to identify these problems in human-produced explanations, this should not be meant to imply that replication of similar problems in explainable AI would be acceptable, nor that exceeding human-level performance is an appropriate threshold for using explainable AI systems." Rewrite the reference to human-level performance serving as a benchmark to AI systems in line 548-549 to read "and to provide a possible baseline, but not threshold of acceptability, for explainable AI systems,..." Rewrite the reference to benchmark on line 561-562 to read: "Humans as a comparison group for explainable AI can partially inform the development of benchmark metrics for explainable AI systems, though alternate metrics representing more useful higher standards should also be developed (https://uwspace.uwaterloo.ca/bitstream/handle/10012/15922/Lin_ZhongQiu.pdf); …" Rewrite the reference to benchmark on line 700 to read "This provides a baseline benchmark, but not an acceptable threshold or exclusive standard, with which to compare AI systems."