**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 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | | **The purpose and scope of this document is unclear.**<br><br>Overall, the paper could use clarifications to the motivation, scope, structure, and goals throughout. A few questions to consider in the next revisions.<br><br>**How do you expect this paper to be used?** Will it help people characterize whether their system is explainable or are you simply articulating some of the information in the field? If the primary purpose is to get people using the same language, then state that and clarify the language in your principles.<br><br>**What type of AI-enabled systems are within the scope of this paper?** Many of these details of the AI systems seem to be implicit. Articulating this more clearly would help readers understand your position better. For example, systems that interact with the world and give complex outputs seem out of the scope. | We suggest that the authors clarify the paper's motivation, scope, structure, and goals throughout. |
| 2 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 6 | **How should Section 6 be used by readers?**<br><br>Section 6 states that humans should be used a "baseline comparison" for explainable AI and provides a brief literature review of how human measure up against the 4 principles. However, the work does not articulate why this baseline would be useful and does not point towards any operational metrics to quantify the comparison.<br><br>Additionally, the literature review is limited to human weaknesses and conflates research about different user populations (e.g. novices vs experts.) Finally, many AI-enabled systems are not intended to mimic human decision making, so using human explanations as a baseline may not inherently be useful. | We suggest cutting this section, as it (1) is unclear how it should be used by the reader, (2) is not justified why humans would make valuable benchmarks, and (3) feels out of place with the rest of the document.<br><br>If the authors keep Section 6, we suggest focusing on humans trained and tasked to give explanations about complex topics (e.g. teachers, journalists, science communication researchers, etc.) as a more appropriate population for benchmarking. Additionally, any benchmark discussion should include human's strengths in addition to their weaknesses. Finally, this section should include ways to operationalize benchmarks for comparison. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | | **Disconnect between "meaningful" description in Section 2 (different user groups) and descriptions of explainable algorithms in Section 5 (model metrics).**<br><br>There is a disconnect between how explainable AI will be leveraged in the real world (Section 2) and the discussion of the current research state of explainable metrics and methods (Section 5). The connection between these sections needs to be made clearer (e.g., how do the metrics described in Section 5 relate to the user groups in Section 2?)<br><br>Additionally, Section 5 describes some models (e.g. regressions) as being "Self-Explainable" because the variable importance can be extracted. However, it is unclear if these "Self-Explainable" models would meet the authors' "meaningful" principle. Consider a regression with thousands of variables/features—is it helpful to have an algorithm with high explanation accuracy if it's not translatable back into human understanding? | Discuss how the descriptions in Section 2 relate to later descriptions in Section 5. Clarify the purpose of Section 5 and how the various proposed methods relate to NIST principles. Finally, if the authors feel that some models are "Self-Explainable", then they should walk through how those model types stand up against the 4 principles. |
| 4 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | | **The AI Output definition seems limited.**<br><br>The scope of AI-enabled systems that the framework applies to is unclear. While the call for comments states the principles aim to represent the "fundamental properties" of explainable AI, various definitions and discussions do not apply to the range of systems discussed in the article.<br><br>A particularly salient point is the discussion of AI-enabled systems of having a sort of simple input-output characterizations (e.g., line 140: "The output is the result of a query to an AI system.") This makes sense for algorithms that output probabilities (predictions), sets of lists (recommendations), and related, enumerable and identifiable output.<br><br>However, this definition does not generalize well to more complex systems such as autonomous cars, which are briefly discussed (c.f., page 4) and presumably within the scope of this paper. Autonomous vehicles have a range of modular capabilities at varying levels of complexity (e.g., the vehicle can trigger wipers; go from point A to point B safely and legally; alert the driver of potential hazards). It is unclear the degree to which the "outputs" produced by the autonomous car are in the scope of this framework. | Clarify the scope of this paper. Which types of AI-enabled systems are within the scope of it? Which, if any, are outside the scope of the paper? How did you evaluate what is in the scope and why are certain things outside of it? |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | | **Assumptions about the algorithm quality should be explicitly stated and/or addressed.**<br><br>One of the primary justifications the author's give for needing Explainable AI is that "…it can be assumed that the failure to articulate the rationale for an answer can affect the level of trust users will grant that system." [lines 125-126]<br><br>However, trust is not desirable when a system is inaccurate and/or unreliable. Explainability should strive to appropriately calibration of users' trust depending on the performance of the system. To that end, Explainable AI must clearly communicate the limitations of the model it is trying to explain. However, the authors seem to make implicit assumptions about model quality that should be addressed more explicitly to scope the paper.<br><br>1) The AI-enabled system is sufficiently accurate to be useful and can be measured. (e.g. lines 214-216)<br>2) AI-enabled decisions are made in discrete events which are obvious to the end user. (e.g. lines 159-163)<br>3) There is a clear-cut line between sensing, detecting, and acting. (e.g. lines 159-163)<br>4) Decisions are made in isolation and do not involve interactions with humans (beyond a single query) or another AI-enabled system. (e.g. lines 159-163)<br>5) Explainable AI-enabled systems are limited to the model itself. (e.g. lines 355 - 358) | If the authors agree with these assumptions, then they should be explicitly stated to properly limit the scope of this document.<br><br>Conversely, if the authors feel that these principles apply to models that violate these assumptions, then the application of these principles to imperfect models should be incorporated in the principle discussions. |
| 6 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 2 | **Make principle titles more detailed and comprehensive.**<br><br>Given that these principles are coming from a first-mover and standardizing organization like NIST, we believe that principles should be comprehensive and clear as possible on a simple read. We suggest updating these phrasings to encompass the full idea on a short read without the need to dig into definitions or the full paragraph. | Update the principle "titles" to be more descriptive. For example.<br>• Explanation -- "Explanations must be supported with evidence and/or reasoning from the system."<br>• Meaningful --"Explanations must be understandable to all relevant stakeholders." |
| 7 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 2.1 | **Naming of the "explanation" principle does not match the content of the definitions.**<br><br>The naming of the principle "Explanation" suggests that it is about giving an explanation; however, all of the principles are related with the explanation: e.g., "Meaningful" principle is about how the explanation should be meaningful.<br><br>Changing Principle 1 to "Supporting evidence", "Evidence production", or something similar is a clearer way to describe what the principle is articulating. | Change the name of the "Explanation" principle to more accurately reflect that it is about the system producing some kind of evidence in its explanation, not just having an explanation. |
| 8 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | | **What content should go in an "explanation"?**<br><br>Section 2.1 states that an explanation should "supply evidence, support, or reasoning…", however there is no discussion of what content should be included in the explanation. Answering the following (non-exhaustive) questions about the inputs, model, and outputs would all contribute to a better understanding of the model's trustworthiness.<br><br>**The Model**: How is the algorithm manipulating the inputs? What are the assumptions of the model?<br><br>**The Inputs**: What data was the model trained on? What data is it currently using? What are the different sources of data. How is the data created? How is data updated? How often? Who is responsible for the data?<br><br>**The Outputs**: How does the output systematically match up against performance and/or discrimination metrics? | There is very little guidance for what the reader should consider in the scope "evidence, support, or reasoning." We suggest incorporating a discussion of what content should be included in an explanation. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 2.2 | **The principle of "Meaningful" should be changed to "user understanding" or something similar.**<br><br>The word "meaningful" is loaded with various interpretations and will likely contribute to inconsistent interpretation of this principle.<br><br>We see a disconnect between "Meaningful" as described as "Systems provide explanations that are understandable to individual users" [line 166] and what is described in section starting on line 181. This section reads as a user acceptance criterion, as stated on line 183, "Generally, this principle is fulfilled if a user can understand the explanation, and/or it is useful to complete a task".<br><br>Given that the focus and operationalization here is on whether different user groups understand certain types of explanations this principle is not well-described as meaningful, i.e., a lay read of meaningful would also include necessarily having what you describe as explanation accuracy so this principle should be renamed. | Given that the operationalization of the "Meaningful" principle is focus on user understanding, it should not be described with this name. We suggest renaming it to something like, "User understanding of explanations" or something similar. |
| 10 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 2.3 | **Evaluation of the "Explanation Accuracy" section.**<br><br>Given that there are still issues inferring any level of "explanation" from a carefully designed and controlled experiment, it is unclear how feasible questions about explanation accuracy are to be solved. Even in models with few variables/features, inferring some level of meaning or importance or variable importance can quickly give different explanations as correlations between variables change (c.f., multi-collinearity).<br><br>Additionally, explanation accuracy in the absence of decision accuracy could result in overconfidence in the system when a surrogate model closely mirrors the underlying model but aligns poorly with reality.<br><br>There are already real-world examples having large [negative] impacts on people's lives. For example, sentencing recommendation algorithms are not trained on any racial information and yet still effectively discriminate based upon race; facial recognition algorithms that give an explanation can wrongfully convict an innocent defendant. These examples also highlight that this principle has important ties to decision accuracy, which should be incorporated and discussed more. | Clarify what kinds of research needs to be done or questions need to be answered regarding explanation accuracy in order to be a useful metric for explainable AI.<br>Add detail about how explanation accuracy is related to and interacts with decision accuracy. |
| 11 | Institute for Defense Analyses (IDA) | Rachel Haga, Brian Vickers | | 2.4 | **Evaluation of the "Knowledge Limits" section.**<br><br>The "knowledge limit" section asserts that there are two types of knowledge limits, but does not provide any literature supporting this taxonomy. If this proposed taxonomy is being introduced by the authors, then more justifications for this classification should be provided. Additionally, the following (non-exhaustive) questions should be considered for Section 2.4.<br><br>• Should the explanation include uncertainty associated with an output?<br>• Should the explanation inform the user if inputs are outside the scope of the training data?<br>• Should the explanation inform the user when an output is an outlier?<br>• Should the explanation inform the user when the model assumptions are violated?<br>• If the designer has placed manual limits on the decision space, should the explanation include when those limits have been reached? | Provide literature supporting the proposed taxonomy, or the author's justification for the novel taxonomy. Additionally, consider expanding the scope of this section. |