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	BlackBerry Corporation		173-174	2.1	*"The Explanation principle obligates AI systems to supply evidence, support, or reasoning for each output."* We respectfully submit our concern that the statement above is too constrictive to be applied to all AI systems across the board. Not all the AI systems are required to supply explanation. Further investigation is required to advance understanding of when and to whom explanation should be provided and for what purpose. In certain cases, supplying explanation could cause unwanted side effects, e.g., exposing vulnerabilities which could be exploited by adversaries. We propose describing the explanation principle in terms of how a system can satisfy the principle.	A system fulfills the explanation principle if it supplies evidence, support, or reasoning for each output.
2	BlackBerry Corporation		183 - 184 392 - 393 411 - 412	2.2 4 5.1	*"Generally, this principle is fulfilled if a user can understand the explanation, and/or it is useful to complete a task"* We generally support the intent of this principle. However, we propose NIST elaborate on why the term "meaningful" is chosen over the term "interpretable". Sections 4 and 5.1 note that the computer science literature often utilizes the term "interpretable"; and that self-explanatory models are often labeled as "interpretable". We hope NIST can make term usage more consistent if the intent is the same or clearly define the differences between the terms "interpretable" and "meaningful". There is an explosion of research effort into explainable AI. To encourage progress and reduce confusion, a clear and concise taxonomy and terminology needs to be established.	

3	BlackBerry Corporation		211 351 – 352 409 - 410	2.3 4 5.1	*“Explanation accuracy is a distinct concept from decision accuracy.”* We generally agree with the intent of the statement above and note that the distinction being called out necessitates it when using the term “accuracy”. We think there may be some ambiguity in references to “accuracy” in Section 4 and 5. Examples are as shown below. We encourage NIST seek to help clarify the ambiguities. *“A key disagreement between philosophies is the relative importance of explanation meaningfulness and accuracy.”* *“Although these simple models are explanations themselves, they are often not always accurate, …”*	
4	BlackBerry Corporation		251-253	3	Referring to the use case of movie recommendations, we note that explanations are not only for the owner’s benefit but also for the user’s. For example, an explanation based on the user’s previous choices may increase his or her willingness to accept a movie recommendation.	User Benefit: This type of explanation is designed to inform a user about an output, persuade a user to accept an output or take a certain action. For example, the explanation could provide the reason for a recommendation in terms of the user’s previous choices, or the reason for a loan application denial. Explanations of this kind contribute to users’ trust.
5	BlackBerry Corporation		289 - 291	3	*“From a practical perspective, explanations can be characterized by the amount of time the consumer of the explanation has to respond to the information and the level of detail in an explanation”.* We understand that the level of detail in an explanation may depend on the amount of time available to the consumer. However, the level of detail does not characterize explanation, meaningfulness (how helpful the explanation is for a consumer to understand it) and explanation accuracy (how precise the explanation reflects the real operation of the model). Further work is required to characterize explanation, meaningfulness and explanation accuracy for types or purpose of explanations.	

6	BlackBerry Corporation		267-270	3	*"System development: This type of explanation assists or facilitates developing, improving, debugging, and maintaining of an AI algorithm or system"* We generally agree with the intent of the statement above. We note substantial research effort into explanatory debugging which supports debugging of leaned programs by an interactive exchange of explanations [ref]. We propose to add the explanatory debugging as an example of explanations for system deployment. [ref] T. Kulesza et al., "Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs," in 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, Leganes, Madrid, Spain, Sep. 2010, pp. 41–48, DOI: 10.1109/VLHCC.2010.15	This category includes the users requiring significant detail and users interacting with the system. For example, this may include the technical staff debugging a vision algorithm with a Gradient-Weighted Class Activation Mapping (GRAD-CAM) based tool [82] and debugging leaned programs by an interactive exchange of explanations [ref].
7	BlackBerry Corporation		724 -726	7	*"To succeed in explainable AI, the community needs to study the interface between humans and AI systems."* We generally agree with the intent of the statement above. We know that human interaction with AI-powered systems can form an extended dialogue or conversation, and that such conversations can be understood from the view of communication and social processes [58]. Some treatments of explainability tacitly model explanations as singular outputs from a lone process, rather than results of an interacting agents. It may be helpful to clarify in this document that the social sciences may be one of the disciplines to draw upon, and particularly that explanations may at least sometimes take the form of extended and interactive explanation processes to encourage broad consideration of the space of explanations.	
8	BlackBerry Corporation		733 - 735	7	*"The common framework and definitions under the four principles facilitate the evolution of explainable AI methods necessary for complex, real-world systems."* We generally agree that a common framework and definition facilitate the evolution of explainable AI methods. A common taxonomy and terminology is a necessary first step to build the common framework and definitions.	