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	IRT Saint-Exupery	DEEL Team	125-144	Introduction	We think the focus is too much articuled around « Trust in the system » or social acceptance. There are other aspects of explanations that should be considered, such as certification for critical systems.	Introduce other scopes of explainable AI in the Introduction.
2	IRT Saint-Exupery	DEEL Team	205-210	Explanation Accuracy	The term « accuracy » is very broad and encompasses very different metrics on explanation methods / systems.	The term « Fidelity » is used in the litterature for « The ability of the explanations to reflect the behaviour of the prediction model. ». There are other properties one may want to achieve such as « stability », « consistency » and « representativity ».
3	IRT Saint-Exupery	DEEL Team	216-217		Measuring a system's accuracy is often quite simple, but measuring the « accuracy » of explanations is much more difficult. The topic of metrics for explainability is growing faster and faster. It would be great to add some recent references.	Emphasize the complexity of computing the « accuracy » of explanations and the various ways of defining « accuracy », and maybe the fact that this is an emerging field of study. Includes more recent references. For this comment and the one above, we propose a (non-exhaustive) list of recent references on the subject. **We propose a list of references in a separate page of this document.**
4	IRT Saint-Exupery	DEEL Team	229-244	Knowledge Limits	Knowledge limits is usually studied outside of the field of Explainability. While we think having a section dedicated to it here is a good idea, we feel that the section is too much isolated compared to the other ones. In particular, there is nothing related to knowledge limits in the « Overview of Explainable AI Algorithms » section. Furthermore, there is more than one way a system can get out of its knowledge limits, not all of them being relevant for explainable AI, but this should be mentioned.	
5	IRT Saint-Exupery	DEEL Team			The term « confidence » has a specific meaning in the machine learning community and should be used carefully to avoid misunderstanding. It is possible to create explanations based on confidence (REF) but using confidence to exclude decisions is usually not a good idea if the confidence is computed a-posteriori from a model which was not trained to exclude decisions (REF).	(REF @article{DBLP:journals/corr/HendrycksG16c, author ={Dan Hendrycks and Kevin Gimpel}, title ={A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks}, journal ={CoRR}, volume ={abs/1610.02136}, year ={2016} })
6	IRT Saint-Exupery	DEEL Team			If the confidence is computed a-posteriori using a third-party system, this third-party system must also be explained.	
7	IRT Saint-Exupery	DEEL Team	245	Types of Explanation	We find the title of this section misleading. «Types of Explanations » might refer to the kind of explanations one produces: attribution maps, linear approximations, counter-examples, etc.	Change the title to « Categories of Explanations » or «Purposes of Explanations ».
8	IRT Saint-Exupery	DEEL Team			This section associates targeted audiences with types of explanations, but there is no introduction to the various existing types of explanations: attribution maps, (local) linear approximation, counter-examples, prototypes, etc.	Add (in this section or somewhere before), a section on the existing kinds of explanations, even if it is not exhaustive, to give readers a sense of what are considered explanations in the current litterature.
9	IRT Saint-Exupery	DEEL Team			We think there is a missing « dimension » in the document regarding the impact of explainability in system design. Which type of explanations are available depends on the workflow used to design the system. E.g., for certifiable systems, one may assume that interpretability or explainability has to be part of the workflow from the beginning, while for « Owner Benefits » use cases, explanations can be provided post-learning, although introducing explainability into the design workflow might improve the explanations.	
10	IRT Saint-Exupery	DEEL Team	355-367	Overview of Explainable AI Algorithms	This section presents multiple points of view from the litterature, which are not all aligned. It should be made clear that there is not a single consensus in the field about explanation types, when to use them, what is explainable, ...	Clearly indicate, here or somewhere else, that there are distinct opinions in the litterature and that a consensus is yet to be reached regarding the questions in this document.
11	IRT Saint-Exupery	DEEL Team		Overview of Explainable AI Algorithms	There is a strong emphasis on transparent models in this section (compared with non transparent models), and the scope of these « transparent » models is not clear (self-explainable vs. transparent).	Use the term « self-explainable » models (as it is used afterwards in the document) which englobes both transparent models (linear models, decision trees, rules list, ...) and explainable-by-design models (disentangled VAEs, prototypes, ...).

12	IRT Saint-Exupery	DEEL Team	405-406	Overview of Explainable AI Algorithms	This sentence implies that the output of a neural network trained in a standard way corresponds to the confidence of the network in the decision. This is not true, unless a specific training procedure has been used.	Remove the sentence or change it to say that confidence can be computed from the output (e.g. from the softmax), but that the output (softmax) is not the confidence. (REF @article{DBLP:journals/corr/HendrycksG16c, author ={Dan Hendrycks and Kevin Gimpel}, title ={A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks}, journal ={CoRR}, volume ={abs/1610.02136}, year ={2016} })
13	IRT Saint-Exupery	DEEL Team	457-461	Self-Explainable Models	Similar to the comment above, this section feels out of place. There is a distinction to be made between transparent models and explainable-by-design models.	Split the « self-explainable » section in two. A first subsection related to transparent models (linear model, decision tree, rules list), and a second subsection related to model that are not transparent but designed to provide explanations or be explainable (disentangled VAEs, prototypes, …).
14	IRT Saint-Exupery	DEEL Team	524-544	Adversarial Attacks on Explainability	This section feels out-of-place. This is directly linked to the « Explanation Accuracy » principle and we feel this would require a whole section.	
15	IRT Saint-Exupery	DEEL Team	524-544	Adversarial Attacks on Explainability	Measuring the « Accuracy » (fidelity, stability, consistency) of explanation methods is probably the biggest challenge in the explainable AI domain, and one that received very little attention in the past. This topic should be given more credit in the document.	
16	IRT Saint-Exupery	DEEL Team	701-702	Discussion and Conclusions	The conclusion mentioned « accountability » of AI systems. There is a missing section in the conclusion about the « accountability » regarding explanations, in particular if the explanation is provided by a different entity than the one providing the AI system.	
17	IRT Saint-Exupery	DEEL Team		Discussion and Conclusions	There should be a clear statement in the conclusion regarding the missing consensus in the domain of explainable AI. A lot of work has been and is currently being done, but even the definition of explanation is subject to divergence.	