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		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		4	Explanation, as a pillar, has some nuanced that is missed surrounding **explanation quality** (e.g. accuracy of explanations), **capturing causal influence in local explanations,** and **preserving privacy**.	1. Explanations must be accurate; and in particular, Shapley Value estimations. For references, see Section V of Datta, Sen, Zick 2016, as well as Sundararajan & Najmi 2020 for additional nuances. 2. Explanations should ideally capture causal influence in local explanations, i.e. by identifying features that are truly driving the model's predictions and teasing them apart from associated features. 3. Explanations should retain privacy, in that we do not disclose sensitive information about individuals when justifying a prediction. See Datta, Sen, Zick 2016 for privacy-preserving explanations for Shapley Value feature importances. This is also a challenge for counterfactual explanations and actionable recourse Karimi et al 2020.
2		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)	368, 462		The provided **definition of global explanations is too narrow**-- it is more than the ability to produce a model that explains/approximates the underlying model.	Global explanations like the examples cited in the paper (SHAP, TCAV, ICE, PDPs) and also other work (see here) provide general visualizations or metrics that characterize model drivers overall or in segments. A lot of the same quality requirements that are necessary for per-decision examples also apply global explanations such as: 1) being causally relevant to the behavior of the model, 2) providing per feature explanations. It is also important for global and per-decision interpretability methods to be consistent. For example, it should not be the case that a feature that is globally important is unimportant for any particular decision.
3		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		2.2	The meaningful pillar should **incorporate a notion of sufficiency.** Explanations must be understandable but also sensical and thorough enough to justify a model's prediction.	We must leverage important input and internal factors as a way to evaluate sufficiency of explanations. Some literature that discusses this: [Leino et al 2018, Wang et al. CVPR Workshop 2020, Lu et al. ACL 2020].
4		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)	134	1	**Stability** is another core component that characterizes trust in AI systems.	People and data change all the time, and models must be robust to this or change alongside them. The Federal Reserve's SR-11-7 memo discusses how model stability and monitoring is integral to model risk management.
5		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		5	Non black-box models like neural networks will **often do better than their "whitebox" counterparts.** It's not entirely accurate that whitebox models are ideal for trustworthy AI– while they might be more interpretable, they also might have poor *knowledge limits*. As an example, deep learning models are high-dimensional functions and their expressive power can capture nuances in data that could make the model robust.	1. It's not just accuracy vs. interpretability but accuracy vs. fairness vs. stability vs. interpretability, etc. In Section 5.1, Shapley values are a case where you can get accurate model explanations even if the model is a blackbox.
6						2. There is a **gradient of black box to whitebox models,** so it is not a binary classification. As an example, in Section 5.1, GA2M is not a true whitebox model. While GA2M contains pairwise interactions that are heatmaps, there feature interactions are hard to interpret or understand.
7		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)	372	5	When discussing counterfactual explanations, it's also worth mentioning **actionable recourse.**	The benefit of counterfactual explanations are that one can stress-test models on modified data points. This ties in directly to the psychology of using trustworthy AI by playing out what-if scenarios and allowing laypeople to understand how decisions can be changed. Some citations: [Rawal & Lakkaraju, 2020; Karimi et. al 2020; Poyiadzi et. al 2020, etc.]
8		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		6.2, 6.3	It is not enough for humans to be able to use model explanations to justify a model decision. Instead, given access to model explanations, humans should be able to **replicate the model decision** (come to the same conclusion on their own).	The report has a lot of great detail on how humans can retroactively justify their own decisions, and thus implicitly trust a model or reverse engineer an explanation from an answer. However, this is dangerous-- and ties into the explanation sufficiency point mentioned above. Some literature that discusses this: [Leino et al 2018, Wang et al. CVPR Workshop 2020, Lu et al. ACL 2020].
9		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		6.4	**Bias and fairness should be mentioned more within "knowledge limits" of a model. Knowing a human's limits involves acknowledging conscious and unsconsious biases that we hold and try to actively compensate for when making a decision. Models should do the same.**	Bias and fairness extends beyond understanding model confidence and identifying out-of-distribution points. Models could confidently reflect human biases in a model that is trained on a lot of data. As humans, we have predefined protected classes that we cannot be discriminated upon. Quantifying disparate impact should be a prerequisite to adoption of any AI, as per [Datta, Sen, Zick 2016; Feldman et al 2016; Dutta et al 2020]
10		Anupam Datta (CMU/Truera), Shayak Sen (Truera), Divya Gopinath (Truera)		6.4	**When discussing knowledge limits and metacognition, it is worth mentioning calibration.**	In section 6.4, the point brought up regarding metacognition has been studied in the statistics and machine learning literature as calibration [Niculescu-Mizil & Caruana, Jiang et al. etc].