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	Duke	Cynthia Rudin	all	all	to be seriously problematic in many ways. I have tried to be constructive to help the authors improve it.... I think the fundamental problem is the same one that characterizes most of the "Explainable AI" literature - it presumes that one would be using a black box and using some mechanism to explain it, rather than starting by default with a model that is inherently interpretable. This problem exists throughout the text, where "explanation methods" are really given highest priority, and interpretability methods are described more as an afterthought. Interpretable machine learning has a much longer history than "explainable" AI, and there is no reason it should be considered as an afterthought, particularly for	

	Duke	Cynthia Rudin	all	all	The paper fails to acknowledges that there are really two fundamentally different types of problems, one type where complex black box models don't help (mainly a mix of meaningful continuous and categorical variables) and problems like computer vision, which are entirely different. I have more discussion of this in Rudin, 2019. Distinguishing between these classes of problems is important because we would never want people to use a complex model when a sparse decision tree would suffice.	acknowledge what problems the paper is referring to with different types of explanations.
	Duke	Cynthia Rudin	mentions of LIME, Grad-CAM, etc. sevderal places throughout	several	The numerous mentions of GradCAM and very little in the way of citations to papers on interpretable neural networks illustrate the point above.	There are numerous papers on truly interpretable neural networks, such as ProroPNet https://arxiv.org/abs/1806.10574 which do not lose accuracy over black boxes.Those methods are more valuable than explainations of black boxes, because their explanations are faithful to the underlying decision-making process of the model.
	Duke	Cynthia Rudin	324		The paper writes: "Rudin [77] and Rudin and Radin [78] argue that models for high-stakes decision must provide explanations that reveal their inner workings. They claim that deep neural networks 326 are inherently black-boxes and should be avoided for high-stakes decisions." This is absolutely NOT what these papers say! Instead they suggest using interpretable deep neural networks for computer vision problems.	The correct wording might be "Rudin [77] and Rudin and Radin [78] argue that it should not be assumed that interpretability must be sacrificed for state-of-the-art accuracy. They provide examples even for deep neural networks in computer vision where interpretable models show no sacrifice in accuracy over black box deep learning methods, and suggest that it is possible that one never needs to sacrifice accuracy for interpretability in high-stakes decisions. They suggest that for high stakes decisions, one should never accept a black box model (even with explanations) unless it can be proven that no interpretable model exists for the same problem with the same level of accuracy. "

	Duke	Cynthia Rudin	336		It states: "In their survey, Gilpin et al. [22] take a similar stance to Rudin [77] and Rudin and Radin [78] in their set of “foundational concepts” for explainability." - NO this is absolutely not true.	I suggest removing the citation to Gilpin et al. as it does not sufficiently survey historical literature on interpretability. It reviewed only a biased selection of recent papers at the time it was published. (A disclaimer: the authors were very junior when this was published so wouldn't be expected to know the field very well.)	
	Duke	Cynthia Rudin			self-explainable models <- "interpretable models," please use the correct historical terminology here	fix terminology (see earlier comments)	removing the citation to Gilpin et al. as it does not sufficiently survey historical literature on interpreta bility. It reviewed only a biased selection of recent papers at the time it was
	Duke	Cynthia Rudin	410		"they are often not always accurate, especially if used without much pre-processing"<- this comment ignores the recent literature on optimized interpretable models. If one uses the 1984 algorithm CART, yes, it will lose accuracy to boosted decision trees. But that isn't a fair comparison. Unfortunately it's the one that almost all "explainability" papers make.	Please see Rudin [19] for a more realistic perspective on this. For most datasets, decision trees perform just fine. The most recent optimal decision tree work is the GOSDT algorithm (Lin et al., 2020) and DL8.5 (Nijssen et al., 2020). References to all recent literature on decision trees is here (https://arxiv.org/abs/2006.08690)	

	Duke	Cynthia Rudin	419		"with the belief that no such trade-off exists for high-stakes decisions." <- there was evidence backing this claim up! Years and years of work.	"With evidence showing that no such tradeoff exists ..."
	Duke	Cynthia Rudin	422		"Lakkaraju and Rudin [50] produces decision lists with improved accuracy." <-- actually, this paper showed how to incorporate costs into decision making. It wasn't arguing about accuracy of predictions, it was showing a method of making interpretable cost-aware decisions.	reframe? Also the latest on decision lists is the CORELS algorithm (Angelino et al 2017).
	Duke	Cynthia Rudin	431		"Bertsimas and Dunn [5] produce a variant of decision trees, called optimal classification trees, that split on mixed integer constraints involving multiple variables. These trees focus on preserving the meaningfulness of decision trees but greatly improving their classification accuracy. " <-- this paper shouldn't be cited, because they didn't make the code publicly available and their results not only weren't very good, but they are not reproducible. The latest papers on decision trees are GOSDT and DL8.5. Bertsimas and Dunn's paper was used to start a company, and requires a license for CPLEX or Gurobi, whereas GOSDT and DL8.5 are free and open.	fix citations
	Duke	Cynthia Rudin	references to SHAP		Shap - Why not look at model reliance as feature importance? That's the one that was traditionally used by Breiman for random forests.	Suggest to include classical notions of variable importance, such as model reliance. Useful references on this topic are here: http://www.jmlr.org/papers/volume20/18-760/18-760.pdf

	Duke	Cynthia Rudin	mentions of LIME		One should probably note that LIME (while a nice idea) has been said many times by many individuals to yield misleading explanations. One should be very careful in using this type of method - only for low-stakes decisions	Suggest to add a warning on uses of approximation models as explanations. They are not "explanations", they are "approximations". An example of where approximations go wrong is in ProPublica's accusation of racial bias to COMPAS. ProPublica approximated COMPAS using a linear model, which they found depended on race, and accused COMPAS of racial bias, even taking into account age and criminal history. However, COMPAS appears to be nonlinear (https://hdsr.mitpress.mit.edu/pub/7z10o269/release/3) so ProPublica's reasoning was invalid. This shows the danger of creating an approximation of a black box and assuming that the black box depends on the same variables! It definitely does not need to!