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	Commente	Paper Line # (if	Paper Section (if	Comment (Include rationale for comment)	Suggested change
1	PwC		125, 133	Intro	In line 125 the authors talk about explainable systems and in line 133 they jump to Explainable AI being one of many properties required of trust in AI systems. The underlying assumption of the document is that one can clearly identify a 'system' as an 'AI system' and therefore require explainability to be one property. Increasingly, AI is being embedded in many devices that may not be easily recognized as an "AI system". For example, a camera auto-adjusting the background light based on the lighting. The software may use AI and it is not clear from the document if the camera would be considered an "AI system" and therefore subject to an explanation requirement. Associated issues include who decides what is an 'AI system' or what is not an AI system.	
2	PwC		159, 160	Four principles of AI	These two lines highlight the requirement for defining an 'output' and the explanation is on the output. Although the explanation is on the 'output' it is relative to the 'input' that the AI system receives and also the 'scope' of the system that is making the decision. The document addresses the 'scope' by having the condition on 'knowledge limits', but does not address the 'input' that goes into the AI system. For example, a traditional thermostat will give a temperature reading within the room where it is placed. It might have a large error band compared to an AI-based thermostat that is continuously receiving input from outside and making adjustments to the temperature readings. The AI-based thermostats will vary in their sophistication based on their input and their 'knowledge limits'. In summary, any explanation of 'output' should be relative to both the 'knowledge limit' and the 'input' it receives.	

3	PwC		289-312	Owner Benefit	This section considers “the amount of time the consumer of the explanation has to respond to the information and the level of detail in an explanation”. While these two dimensions are important, they are not the only ones. Another key dimension of an explanation is the “prior knowledge” of the consumer of the explanation. The better the prior knowledge - the less detailed the explanation needs to be. While this may be covered somewhat by the five categories (Lines 251-276) of users or explanations - they need to be further qualified that these explanations would need to vary by the ‘prior knowledge’ of the user.	
4	PwC			Explanation	requires explanation in all instances, does not consider if an explanation is actually required. The look at time to process vs level of detail but this should be more of a risk-based evaluation. This would also make it more consistent with what the EU Commission is stating. It should also consider the consumer need -- considering the risk of harm (criticality) as well as the hurdle or burden for a user to trust the decision (vulnerability). High vulnerability would be for highly traimned professionals working in critical areas e.g. pilots, doctors.	
5	PwC		190	Meaningful	“Affected party” as an explicit group to be considered	
6	PwC		217	Accuracy	Many explanation methods are rough approximations (e.g surrogate models, SHAP values). In practice this is a hard principle to comply with. The accuracy can be difficult to measure, and it can constantly be changing, especially as many models are retrained constantly or with high frequency (e.g. Amazon deploys potentially thousands of models a day). This may lead to additional uncertainty which does nothing to build trust.	
7	PwC		234	Knowledge Limits	Should also be considered in the context of the risk of the application itself. There are going to be times when users do not want to know how uncertain a model is, and may not be able to understand what that means (think of a self driving car… I do not think a passenger will be happy if the car starts saying that it is 65% sure it sees a pedestrian, htough in thepry high safety critical applications should fall into the range of a requirement). These limits may be more practical in the overall process rather than the model itself. In the self driving car example that may be a popup that suggests the user takes control over the vehicle because conditions have worsened.	
8	PwC		368		Global and local (per decision) explanations are introduced here. They should be considered in the principles above based on what users and other stakeholders need of their systems	

			285		Surrogate models are introduced but never defined nor given examples of. These are some of the more practical explanation mechanisms in development	
				General comment	Need to distinguish explanation in the context of model development vs deployment.	
			531	Adversarial Attacks on Explainability	Introduce fairness but never define it. There is no evidence that a more interpretable model is inherently more fair, and how an ensemble would hide the unfairness of an underlying model. There are many different ways to measure fairness (see Fairness Definitions Explained - Verma et al 2018: http://www.ece.ubc.ca/~mjulia/publications/Fairness_Definitions_Explained_2018.pdf)	
			548	Humans as a Comparison Group for Explainable AI	A system may not just be a human and a model. Often times models are latent in other processes, without humans in the loop. For instance a whole system for processing invoices would have OCR to convert PDF to readable format, an information extraction model to pull information from the invoice into a spreadsheet, and maybe some kind of RPA which tabulates and issues a payment. A human may just be involved with supplying the invoice and doing some amount of QA at the end.	