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
		John S. Hyatt	181-244	2	As written, the paper apparently does not allow for the possibility that a system could produce quantified/calibrated uncertain outputs. For example, in some cases it might be valuable to know that there is a 50% likelihood of prediction A, 25% likelihood of B, 5% each of C-G, and <1% of H+. This can be particularly important in risk evaluation, when a human or another system uses the output to inform a downstream decision making process, etc. (I phrase this around classifiers, but it's equally applicable to all kinds of AI/ML systems.) Uncertainty, calibration, etc. are all ongoing areas of research in explainable AI. Obviously, sometimes it's only acceptable to have a yes/no prediction, with a high certainty requirement; however, uncertain partial predictions can still be more useful than no prediction at all, provided the predicted uncertainty is a good estimate of the true uncertainty. Humans have to take calculated risks based on imperfect knowledge all the time, and AI system predictions should allow that, where appropriate, by providing quantitative estimates of the predicted uncertainty even when the main prediction is highly uncertain.	I think a sentence or two to this effect could be added to the Knowledge Limits principle without losing anything or overly complicating the issue. For example, changing line 169 to: "The system only operates under conditions for which it was designed. Additionally, it either (i) only operates when it has sufficient confidence in its output, or (ii) produces a well-calibrated uncertainty estimate for its output that itself meets the four principles listed here." There would be some supporting changes later. For example, in section 2.4, a reasonable addition might be (at the end): "I am sure this is a bird. The image is too blurry to identify the bird, but based on coloration it might be one of these species with such-and-such probability, and it is definitely not any of these other species."