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	Ethical AI Consortium, Inc.	Sarah Alt	128	1. Introduction	There are many instances where people don't have a choice that AI is used on them. They don't have a choice to accept or adopt.	Suggest also acknowledging there are real instances where acceptance and adoption aren't even an option.
2	Ethical AI Consortium, Inc.	Sarah Alt	156	2. Four Principles of Explainable AI	I believe the word is "regulator" is supposed to be "regulatory"?	Change to "regulatory"?
3	Ethical AI Consortium, Inc.	Sarah Alt	181	2.2 Meaningful	I'm sure there was much discussion and debate on whether to use "Understandable" instead of "Meaningful". I feel something doesn't have to be meaningful (having signficance or purpose) that is still understandable (comprehendable). I don't believe we would require Explainable AI to be meaningful. In fact, there may be many rather frivolous applications of AI, depending upon one's viewpoint. But we could still explain them in an understandable way.	Suggest changing this to "Understandable"; and all places where some version of meaningful as the root or implied definition is used (e.g. line 199 "meaninfulness").
4	Ethical AI Consortium, Inc.	Sarah Alt	208-209	2.3 Explanation Accuracy	This phrase: "...an explanation that correctly reflects a system's process for generating its output" is a great and precise definition of what "Accuracy" means.	Make it really simple and state just that. "We define Explanation Accuracy as an explanation that correctly reflects a system's process for generating its output."
5	Ethical AI Consortium, Inc.	Sarah Alt	209-210	2.3 Explanation Accuracy	This is a confusing sentence: "The Explanation Accuracy principle imposes accuracy on a system's explanation." - It is also a circular definition.	If we succinctly state what I suggest in comment #4, you don't need this sentence at all.
6	Ethical AI Consortium, Inc.	Sarah Alt	211-217	2.3 Explanation Accuracy	While informative, I don't think you need this paragraph.	Suggest dropping this paragraph.
7	Ethical AI Consortium, Inc.	Sarah Alt	225	2.3 Explanation Accuracy	Remove "more than one type of of [sic] explanation"	Suggest replacing it with "various levels of detail in its explanations tailored and relevant for various audiences."
8	Ethical AI Consortium, Inc.	Sarah Alt	228	2.4 Knowledge Limits	Generally speaking, do we also mean it is used for its intended purpose(s) only?	If so, suggest we address this someplace in this definition too.
9	Ethical AI Consortium, Inc.	Sarah Alt	231	2.4 Knowledge Limits	This is confusing: "...or their answers are not reliable."	Do we mean perhaps: "...or cases where their answers are not reliable."?
10	Ethical AI Consortium, Inc.	Sarah Alt	237-239	2.4 Knowledge Limits	This set of sentences is confusing..."The system could return an answer to indicate that it could not find any birds in the input image; therefore, the system cannot provide an answer. This is both an answer and an explanation." - We say the system could return an answer in the same sentence we say it cannot provide an answer, and in the very next sentence we say it is an answer.	Not sure how to suggest we fix this.
11	Ethical AI Consortium, Inc.	Sarah Alt	235-244	2.4 Knowledge Limits	I believe the document should always use human examples, rather than a birds example. While it is a "safe" example, it waters down why and in what instances we remain most concerned about explainability, which is where fundamental human rights, discrimination and bias are involved.	Suggest replacing this with a human example.
12	Ethical AI Consortium, Inc.	Sarah Alt	250	3. Types of Explanations		Suggest adding "The categories described below were not designed to be exhaustive, nor exclusive, meaning a given individual could be more than one audience member in a given AI system situation."

13	Ethical AI Consortium, Inc.	Sarah Alt	251-276	3. Types of Explanations	These categories/groups feel confusing. Especially "societal acceptance" doesn't feel like a group of people as much as it is a concept that all explanations should be aiming to achieve. These feel like a combination of groups of professions/people and concepts.	Suggest using the following 5 groups: Builders/Developers, Regulatory/Legal/Compliance, Buyers/Subscribers, Users, Beneficiaries
14	Ethical AI Consortium, Inc.	Sarah Alt	256	3. Types of Explanations	If you keep this section, change "users" to "people". Some people in society aren't actually the users of the AI. It is used on them, which means they are intended to be the beneficiaries - whether they had a choice or not.	
15	Ethical AI Consortium, Inc.	Sarah Alt	259-264	3. Types of Explanations	Keep this group, but drop using the word "user".	Change this to "a person", "an individual", or "someone". We still need to distinguish that there are people who use AI and there are people on whom AI is used.
16	Ethical AI Consortium, Inc.	Sarah Alt	267	3. Types of Explanations	Change "includes" to "include"	
17	Ethical AI Consortium, Inc.	Sarah Alt	271-276	3. Types of Explanations	I don't understand how "operator of a system" is different than "user" (line 251). Likewise, if we mean "operator", then why do we call it "owner"? There are many instances where an operator of a system is not the owner. Similarly, the example provided confuses me more. I don't "own" my streaming movie services. I am a Subscriber or Beneficiary of the AI system a developer developed or that Netflix decides to use with my data. I'm definitely not an owner. But I might be an operator?	See Comment #13 for a suggested nomenclature to clear this up.
18	Ethical AI Consortium, Inc.	Sarah Alt	295-301	3. Types of Explanations	This is a good human example! Nice job.	
19	Ethical AI Consortium, Inc.	Sarah Alt	322-323	4. Overview of principles in the literature	Change this from a bird example.	Follow the human example if you decide to change it.
20	Ethical AI Consortium, Inc.	Sarah Alt	523	5.4 Adversarial Attacks on Explainability	This seems a bit buried in this section and awkwardly placed. It also doesn't seem like this would be the only caution or risk we would want to point out.	Does it belong here at all? Is this another paper all together?
21	Ethical AI Consortium, Inc.	Sarah Alt	590	6.2 Meaningful	This first sentence should be carefully worded to match the same definition used for "Meaningful" (or "Understandable") as used in the beginning of the paper. Here we introduce slightly different words to describe "Meaningful". If we are making a comparison between the capabilities of AI systems to explain "themselves" and humans to explain them, we should hold to the same controlled definition of the standards.	
22	Ethical AI Consortium, Inc.	Sarah Alt	605-609	6.2 Meaningful	These same issues exist with AI systems explaining "themselves" too.	
23	Ethical AI Consortium, Inc.	Sarah Alt	611-612	6.3 Explanation Accuracy	Again, we use a different definition here than earlier in the paper. Isn't "accuracy" simply: does the explanation factually and correctly describe what the AI system does?	This may be intentional and rational, so disregard if there is a good explanation for it.

24	Ethical AI Consortium, Inc.	Sarah Alt	697-707	7. Discussions and Conclusions	The subtitle of the posting for this paper is: "Technical agency proposes four fundamental principles for judging how well AI decisions can be explained." Most readers would expect to find these in the paper. This section contains perhaps the most important statements that should not be buried in my opinion. If NIST is suggesting that explainability is achieved by both system and human involvement, then let's just say that at the beginning. Provide the four guiding principles and state the role that both the AI system (as designed by humans) and the humans themselves play in achieving explainability. Perhaps some of this paper belongs in separate rationale papers to explain why or how we arrived there, but the core of the guidelines should be ready to use.	Good start and maybe we need a more consumable piece that results from this for lay people to understand. You may already be planning to do so.