Introducing the framework

1. **Can AI systems be classified <u>consistently & reliably</u> with the core criteria?**

I think overall this provides the ability to classify AI, but only in a limited way, for example I think the user should explain more about the types of models used (noted later in the text). Also there is not enough to determine if the AI is "responsible" or not. For example data privacy, environmental impact, and bias go largely undiscussed/unreported in the sample use cases.

I. CLASSIFICATION FRAMEWORK

1) CONTEXT

- At which operational model do you have to evaluate model performance on various communities such as monitoring for bias? Here at the operational level I would assume at least.
- Does system monitoring have to be automated or is manual monitoring considered ok? I would suggest that in TRL9 you have automated monitoring/continuous learning systems and in TRL8 it is at least manual monitoring. Putting the entire process of monitoring (which is a huge area of development) into one TRL does not seem appropriate.
- I think this should be "Stakeholder impacted by the system may include:" to emphasize that this is by no means a complete list of possible stakeholders

2) DATA AND INPUT

- *Synthetic data:* Does this include data generated by reinforcement learning?
- '*Identifiability' of personal data*: Is there going to be no discussion of data protection and security? It seems like a fairly large area to avoid unless that is covered by some related rules.

3) AI MODEL

- Perhaps another question, "Is there a way to define the decision space of the AI system?"
- **"Which criteria should be in a more detailed, technically-oriented framework?** " This seems to need more detail rather than just listing one of these three types. Ideally knowing the purpose of model (anomaly detector) and the mode used (decision tree) I think this need to be more technical.
- Model inference is pretty important, this should be a required criteria.
- Actually building the model is not included in this list of two items? You could make that clearer by saying "include but not limited to"

4) TASK AND OUTPUT

- *Recognition*: I assume this also includes things like image segmentation and object detection. If that is not the case it is not clear from explanation that those are not included

- We are not going to discuss anywhere in regards to the impact on the planet of training the models (e.g deep learning models)? That is fine if we are just worried about AI classification and not the broader context of responsible AI.

II. APPLYING THE FRAMEWORK

- *AI model*: You talked about it a lot in the text, but there is no specific question on how explainable the algorithm is?

- *Model type – hybrid:* This does not seem like enough information about the model to classify it. I would like to see some explanation of the models used (decision trees, neural networks, etc). Just "hybrid" seems insufficient

- *Figure 7. AI system to help manage a manufacturing plant*: This graphic is very helpful and at least lists to components of the AI model. It seems optional though because the other examples do not include it. It would be helpful to know all the models in the AI system to classify it. Eg. No where in the rubric do we state that there is an anomaly detector

-