NIST AI RMF		Requirements/ objectives	Comments	RAII	Requirements/ objectives	RAII questions
Technical		17 Technical characteristics in the AI RMF taxonomy refer to factors that are **under the direct** **18 control of AI system designers and developers, and which may be measured using standard** **19 evaluation criteria.** Technical characteristics include the tradeoff between convergent 20 discriminant validity (whether the data reflects what the user intends to measure and not other 21 things) and statistical reliability (whether the data may be subject to high levels of statistical 22 noise and measurement bias). Validity of AI, especially machine learning (ML) models, can be 23 assessed using technical characteristics. Validity for deployed AI systems is often assessed with 24 ongoing audits or monitoring that confirm that a system behaves as intended. It may be possible 25 to utilize and automate explicit measures based on variations of standard statistical or ML 26 techniques and specify thresholds in requirements. Data generated from experiments that are 27 designed to evaluate system performance also fall into this category and might include tests of 28 causal hypotheses and assessments of robustness to adversarial attack.	I appreciate the distinction between machine issues vs. human issues, as represented in the technical and socio-tehcnical categories of the framework, however, from experiencing with testing these objectives between use cases I'm not sure what value add this distinction brings as there are aspects of technical that are still influenced by how a person designed a machine. Additionally, if the idea is that technical can be evaluated using statistical tests and socio-technical can only be process-based evaluations, I wouldn't agree with that assertion. We have found from testing that it is often difficult to even make decisions between what is being tested. That said, it's good to point out that there could be risks based on system operations vs those where there are more human-based calibrations. My only concern is that by putting them into two categories it makes it seem like there aren't important interdependancies and that there can be distinct evaluations done within each of these indpendent categories. Generally, I think that there needs to be a stronger emphasis on data. Some of the descriptions reference it, but there is more emphasis on the oversight of the model. I think that it would be important to call that out in fairness/ managing bias, which I would include in the technical portion of the framework.			
	Accuracy	**Accuracy indicates the degree to which the ML model is correctly capturing a relationship that** **7 exists within training data.** Analogous to statistical conclusion validity, accuracy is examined via 8 standard ML metrics (e.g., false positive and false negative rates, F1-score, precision, and recall), 9 as well as assessment of model underfit or overfit (high testing errors irrespective of error rates 10 in training). It is widely acknowledged that current ML methods cannot guarantee that the 11 underlying model is capturing a causal relationship. Establishing internal (causal) validity in ML 12 models is an active area of research. AI risk management processes should take into account the 13 potential risks to the enterprise and society if the underlying causal relationship inferred by a 14 model is not valid, calling into question decisions made on the basis of the model. Determining a 15 threshold for accuracy that corresponds with acceptable risk is fundamental to AI risk 16 management and highly context-dependent	In the scope it indicates "25 The NIST AI RMF offers a process for managing risks related to AI systems across a wide 26 spectrum of types, applications, and maturity" so why is there a a reference to ML models. Maybe I missed that AI was being defined as having an ML model. This is a big assumption to make and therefore would not be applicable to many automated predictive systems. Is this the intent?	System operations	1. System scope and function 2. Human-in-the-loop 3. Model is fit for purpose 4. Representative and relevant data 5. Data quality 6. Model accuracy	
	Reliability	**Reliability indicates whether a model consistently generates the same results, within the bounds** **19 of acceptable statistical error.** Techniques designed to mitigate overfitting (e.g., regularization) 20 and to adequately conduct model selection in the face of the bias/variance tradeoff can increase 21 model reliability. The definition of reliability is analogous to construct reliability in the social 22 sciences, albeit without explicit reference to a theoretical construct. Reliability measures may 23 give insight into the risks related to decontextualization, due to the common practice of reusing  10 ML datasets or models in ways that cause them to become disconnected from the social contexts 2 and time periods of their creation. As with accuracy, **reliability provides an evaluation of the** **3 validity of models, and thus can be a factor in determining thresholds for acceptable risk.**				

	Robustness	**5 Robustness is a measure of model sensitivity, indicating whether the model has minimum 6 sensitivity to variations in uncontrollable factors. A robust model will continue to function 7 despite the existence of faults in its components. The performance of the model may be 8 diminished or otherwise altered until the faults are corrected.** Measures of robustness might 9 range from sensitivity of a model's outputs to small changes in its inputs, but might also include 10 error measurements on novel datasets. Robustness contributes to sensitivity analysis in the AI 11 risk management process.	Robustness, Security, and Safety	1. Data drift 2. System acceptance testing 3. Contingency planning 4. Reliability
	Resilience or Security	**3 A model that can withstand adversarial attacks, or more generally, unexpected changes in its 14 environment or use, may be said to be resilient or secure**. This attribute has some relationship to 15 robustness except that it goes beyond the provenance of the data to encompass unexpected or 16 adversarial use of the model or data. Other common ML security concerns relate to the 17 exfiltration of models, training data, or other intellectual property through AI system endpoints.		
Socio- Technical		19 Socio-technical characteristics in the AI RMF taxonomy refer to how AI systems are used and 20 perceived in individual, group, and societal contexts. **This includes mental representations of 21 models, whether the output provided is sufficient to evaluate compliance (transparency), whether 22 model operations can be easily understood (explainability), whether they provide output that can 23 be used to make a meaningful decision (interpretability), and whether the outputs are aligned 24 with societal values.** Socio-technical factors are inextricably tied to human social and 25 organizational behavior, from the datasets used by ML processes and the decisions made by 26 those who build them, to the interactions with the humans who provide the insight and oversight 27 to make such systems actionable. 28 Unlike technical characteristics, **socio-technical characteristics require significant human input 29 and cannot yet be measured through an automated process**. Human judgment must be employed 30 when deciding on the specific metrics and the precise threshold values for these metrics. The 31 connection between human perceptions and interpretations, societal values, and enterprise and 32 societal risk is a key component of the kinds of cultural and organizational factors that will be 33 necessary to properly manage AI risks. Indeed, input from a broad and diverse set of 34 stakeholders is required throughout the AI lifecycle to ensure that risks arising in social contexts 35 are managed appropriately.		

Explainability	**2 Explainability seeks to provide a programmatic, sometimes causal, description of how model** **3 predictions are generated.** Even given all the information required to make a model fully 4 transparent, a human must apply technical expertise if they want to understand how the model 5 works. Explainability refers to the user's perception of how the model works – such as what 6 output may be expected for a given input. Explanation techniques tend to summarize or visualize 7 model behavior or predictions for technical audiences. Explanations can be useful in promoting 8 human learning from machine learning, for addressing transparency requirements, or for 9 debugging issues with AI systems and training data. However, risks due to explainability may 10 arise for many reasons, including, for example, a lack of fidelity or consistency in explanation 11 methodologies, or if humans incorrectly infer a model's operation, or the model is not operating 12 as expected. Risk from lack of explainability may be managed by descriptions of how models 13 work to users' skill levels. Explainable systems can be more easily debugged and monitored, and 14 lend themselves to more thorough documentation, audit, and governance. 15 Explainability is related to transparency. Typically the more opaque a model is, the less it is 16 considered explainable. However, transparency does not guarantee explainability, especially if 17 the user lacks an understanding of ML technical principles	This doesn't include criteria for how a model was trained. In the definition for explainability it speaks to explainability being the underlying mechanisms of an algorithms operation, the input, including the selection criteria is an important feature for understanding explainability of a model. Additionally, understanding that this is under the category of soci-technical with an emphasis on the human intervention, but there are technical ways to interpret explainability. Re 7 - Calling it a user's "perception" of how the model works makes it appear as a subjective approach. Whereas, there are very specific technical means to understand what drives an output based on the inputs. Re: 13 - A very subjective approach i.e. "descriptions". Where explainability fails is when technical approaches "generalize" the explanation of the drivers of the model output. Understanding how the model behavior changes based on the range and permutation of inputs (i.e. different slices/segments of the input space) can vary significantly and can lead to an incorrect understanding of how the model works in specific instances. This is one of the biggest risks/limitations of generalized explainability. We need to be able to understand the variations of the models for different situations and "scale" explainability of models.	Explainability and Interpretability	1. Communication about the outcome 2. Notification 3. Recourse 4. Clear understanding of how the system arrives at a decision or function
Interpretability	**19 Interpretability seeks to fill a meaning deficit. Although explainability and interpretability are** **20 often used interchangeably, explainability refers to a representation of the mechanisms** **21 underlying an algorithm's operation, whereas interpretability refers to the meaning of its output** **22 in the context of its designed functional purpose.** The underlying assumption is that perceptions 23 of risk stem from a lack of ability to make sense of, or contextualize, model output appropriately. 24 **Model interpretability refers to the extent to which a user can determine adherence to this** **25 function and the consequent implications of this output upon other consequential decisions for** **26 that user.** Interpretations are typically contextualized in terms of values and reflect simple, 27 categorical distinctions. For example, a society may value privacy and safety, but individuals 28 may have different determinations of safety thresholds. Risks to interpretability can often be 29 addressed by communicating the interpretation intended by model designers, although this 30 remains an open area of research. The prevalance of different interpretations can be readily 31 measured with psychometric instruments.	This is not a definition for interpretability that we have come across. We would include this more in transparency. All of these would greatly benefit from examples as it would require us to think about what these differences actually are and how they manifest in different systems. That might help with some of the nomenclature challenges. Not to be picky, but just so that we are clear what to evaluate. I think that it's great to point out that different societal values may lead to different configuration of a model. I haven't thought about categorizing this as interpretability previously. We track this in system operations recognizing that there are decisions that need to be made by default by the development team at the model level. However, these trade off decisions should be made as part of a diverse committee review proceess, understood and documented. I would add this more in safety, but I don't think it matters where it is as long as it is covered.		
Privacy	33 **Privacy refers generally to the norms and practices that help to safeguard values such as human** **34 autonomy and dignity.** These norms and practices typically address freedom from intrusion, 35 limiting observation, or individuals' control of facets of their identities (e.g., body, data, 36 reputation). Like safety and security, specific technical features of an AI system may promote 37 privacy, and assessors can identify how the processing of data could create privacy-related 38 problems. However, determinations of likelihood and severity of impact of these problems are 39 contextual and vary among cultures and individuals.			
Safety	2 **Safety as a concept is highly correlated with risk and generally denotes an absence (or** **3 minimization) of failures or conditions that render a system dangerous.** As AI systems interact 4 with humans more directly in factories and on the roads, for example, the safety of these systems 5 is a serious consideration for AI risk management. Safety is often – though not always – 6 considered through a legal lens. Practical approaches for AI safety often relate to rigorous 7 simulation and in-domain testing, real-time monitoring, and the ability to quickly shut down or 8 modify misbehaving systems		Robustness, Security, and Safety	

	Managing bias	10 NIST has identified three major categories of bias in AI: **systemic, computational, and human**. 11 Managing bias in AI systems requires an approach that considers all three categories. 12 Bias exists in many forms, is omnipresent in society, and can become ingrained in the automated 13 systems that help make decisions about our lives. While bias is not always a negative 14 phenomenon, certain biases exhibited in AI models and systems can perpetuate and amplify 15 negative impacts on individuals, organizations, and society, and at a speed and scale far beyond 16 the traditional discriminatory practices that can result from implicit human or systemic biases. 17 Bias is tightly associated with the concepts of transparency and fairness in society. See NIST 18 publication “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.”	Breaking down what systemic, computational and human bias are would be useful additons. A bigger question I suppose though is why just manage bias and not other harms when there is also a fairness section? And then how does this fit with the framing risk section? We strongly agree with the flow of identifying the risk (map), measure against the framework, and then perform ongoing oversight (manage). This could make it redundant or confusing to identify here. We have classified harms in the following ways: Unintentional - Harms arise from AI systems behaving in unintended ways Intentional - Harms arise from adversaries or bad actors purposefully using AI in a malicious way Systemic - Unintended consequences from the deployment of technology that shape the broader environment We are testing all of this right now so the definitions are subject to change, but we wanted to address the fact that there are actions that can be taken that will change the outcome (unintended), that there are often intentional trade-offs as stated above in the interpretability discussion, and systemic being known issues that are difficult to change through the design/ operations of an AI system (eg. an automated lending system using a FICO score).		
Guiding Principle		28 **Guiding principles in the AI RMF taxonomy refer to broader societal norms and values that** **29 indicate societal priorities.** While there is no objective standard for ethical values, as they are 30 grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed 31 that AI technologies should be developed and deployed in ways that meet contextual norms and 32 ethical values. When specified as policy, guiding principles can enable AI stakeholders to form 33 actionable, low-level requirements. Some requirements will be translated into quantitative 34 measures of performance and effectiveness, while some may remain qualitative in nature. 35 Guiding principles that are relevant for AI risk include fairness, accountability, and transparency. 36 Fairness in AI systems includes concerns for equality and equity by addressing socio-technical 37 issues such as bias and discrimination. Individual human operators and their organizations 38 should be answerable and held accountable for the outcomes of AI systems, particularly adverse Initial Draft 13 1 impacts stemming from risks. Absent transparency, users are left to guess about these factors and 2 may make unwarranted and unreliable assumptions regarding model provenance. Transparency 3 is often necessary for actionable redress related to incorrect and adverse AI system outputs.	Not sure that I understand this section, would it be worthwhile to establish Guiding Princinples that the AI RMF is grounded in and then state some objectives/ desired postures for each of these sections? Maybe it's just the order, wouldn't the guiding principles guide the rest of the framework?		
	Fairness	**5 Standards of fairness can be complex and difficult to define because perceptions of fairness** **6 differ among cultures**. For one type of fairness, process fairness, AI developers assume that ML 7 algorithms are inherently fair because the same procedure applies regardless of user. However, 8 this perception has eroded recently as awareness of biased algorithms and biased datasets has 9 increased. Fairness is increasingly related to the existence of a harmful system, i.e., even if 10 demographic parity and other fairness measures are satisfied, sometimes the harm of a system is 11 in its existence. While there are many technical definitions for fairness, determinations of 12 fairness are not generally just a technical exercise. Absence of harmful bias is a necessary 13 condition for fairness.	Is there an objective statement that could be made?	Bias and Fairness	1. Human rights/ ethics acceptance 2. Bias training and education 3. Test for unwanted bias
	Accountability	**15 Determinations of accountability in the AI context are related to expectations for the responsible** **16 party in the event that a risky outcome is realized**. Individual human operators and their 17 organizations should be answerable and held accountable for the outcomes of AI systems, 18 particularly adverse impacts stemming from risks. The relationship between risk and 19 accountability associated with AI and technological systems more broadly differs across cultural, 20 legal, sectoral, and societal contexts. Grounding organizational practices and governing 21 structures for harm reduction, like risk management, can help lead to more accountable systems.	Accountability is also where governance including diverse independent review is typically included, ongoing monitoring of a system, documentation, recourse, notfication? Maybe these are all things that would be outlined in the implementation guide if it's created, but I think that stressing the importance of good governance here and how it relates to the AI RMF is important.	Accountability	1. Clear oversight process for implementation of AI 2. Independent review process and ongoing monitoring

Transparency	**23 Transparency seeks to remedy a common information imbalance between AI system operators** **24 and AI system consumers.** Transparency reflects the extent to which information is available to a 25 user when interacting with an AI system. Its scope spans from design decisions and training data 26 to model training, the structure of the model, its intended use case, how and when deployment 27 decisions were made and by whom, etc. Absent transparency, users are left to guess about these 28 factors and may make unwarranted and unreliable assumptions regarding model provenance. 29 Transparency is often necessary for actionable redress related to incorrect and adverse AI system 30 outputs. A transparent system is not necessarily a fair, privacy-protective, secure, or robust 31 system. However, it is difficult to determine whether an opaque system possesses such 32 desiderata, and to do so over time as complex systems evolve.	Consumer protection	1. Transparency to the use and data subject 2. Harm to individuals/ incident reporting 3. System protects individual's or groups privacy