CALYPSOAI

CalypsoAI's Response to the Request for Comment on NIST's AI Risk Management Framework: Second Draft

September 26, 2022

Artificial intelligence (AI) has the ability to benefit nearly all aspects of the United States' society, economy, and national security; however, AI also poses a unique set of risks. As AI systems are increasingly utilized, managing the associated risks is critical. CalypsoAI is encouraged by the steps the National Institute of Standards and Technology (NIST) has taken to create this AI Risk Management Framework (RMF), which will ensure AI systems in the U.S. are safe, secure, trustworthy, and transparent.

CalypsoAI provided input to both the initial AI RMF request for comments released in 2021 and the request for comments following the release of the AI RMF first draft in early 2022. In this second draft, we are pleased with the emphasis that NIST has placed on the testing, evaluation, verification and validation (TEVV) of AI systems. Combined with the DoD's recent Responsible AI strategy, it is clear that the U.S. government is prioritizing this critical aspect of AI development.

CalypsoAI agrees that standards will continue to evolve with the technology landscape. However, this need not mandate the creation of a cumbersome validation process requiring sign-off from multiple stakeholders each time an entity seeks to deploy AI models, nor cause a delay in establishing a standardized validation method. Given CalypsoAI's expertise in third-party, independent AI/machine learning (ML) model validation, we know it is possible to implement an automated TEVV process that mitigates risk and builds trust. Hence, our comments to this second draft of NIST's AI RMF address AI TEVV: the remaining gaps in the AI RMF, what we are pleased to see included in this second draft, and further suggestions as we move toward a final version.

Remaining gaps in the RMF:

- While TEVV is emphasized in the framework, *independent* TEVV is not specifically called out as a critical element. It is important that a third-party independently conducts TEVV so organizations / developers are not 'checking their own homework.'
- In order to ensure consistency and hold every AI/ML model to the same measure, TEVV should be standardized and repeatable across the enterprise to ensure widespread mission success
- AI/ML model performance must be evaluated *beyond* F1 scores to include testing for operational conditions. The AI RMF mentions the need for deployment context and acknowledges the risks in operational environments are different than laboratory environments. However, it does not mention what general types of operational conditions

to test for (i.e. privacy, robustness against adversarial attacks, weather scenarios). The RMF should clarify what types of tests should be conducted to account for operational environments.

Key Passages from the Second Draft:

AI RMF: TEVV tasks are identified at every stage of the AI lifecycle, including design and planning (validating capabilities relative to the intended context of application); development (pre-deployment model validation and assessment); deployment (system validation, with recalibration based on internal and external factors); and operations (ongoing monitoring and testing).

- ***CalypsoAI:*** TEVV is crucial pre- and post-deployment; however, pre-deployment TEVV is especially important. It allows for the evaluation of model performance, identification of risks associated with deployment, and the opportunity to improve a model before it is put into action, all of which reduce negative consequences. NIST's inclusion of TEVV across the entirety of the AI lifecycle illustrates the positive impact a repeatable, trustworthy AI pipeline can have at all stages and for all stakeholders.

AI RMF: TEVV is "foundational to risk management, providing knowledge and feedback for AI system management and governance." (Pg 6)

- ***CalypsoAI:*** TEVV is foundational to AI risk management; however, it provides knowledge and feedback for *more* than AI system management and governance. In addition to these critical tasks, TEVV establishes context and frames risks related to an AI system, and can provide quantitative measurements for model performance.

AI RMF: A definition for 'trustworthy AI,' characterized as an AI system that is "valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced." (Pg 1)

- ***CalypsoAI:*** We concur with this definition of trustworthy AI; however, we recommend it also include alignment to departmental level policy, governance, doctrine, and standards.

AI RMF: Emphasis on understanding how the AI system performs in its deployment environment to more accurately conduct a risk-benefits tradeoff (thinking "more critically about context and potential or unexpected negative and positive impacts"). This information can then be leveraged to "design, develop, evaluate, and use AI systems with impact in mind and prevent, preempt, detect, mitigate, and manage AI risks." (Pg 2)

- ***CalypsoAI:*** While TEVV is integral to understanding model performance and providing critical information for evaluating risk, automating TEVV both pre- and post-deployment accelerates the process, streamlines data scientists' tasks, and shifts dependence away from arbitrary model evaluation metrics, such as F1 scores, Receiver Operating Characteristic (ROC) Curves, Precision, and Recall. The latter benefit is especially important because the identified metrics only offer insight into how a model performs on its training data, not in its deployment environment. As a high-risk national security example, a model that is used in a UAV over Afghanistan will not perform the same over

Ukraine because of differing ground conditions. Model testing must account for these differences. How does the model deployed on the UAV perform when exposed to foggy conditions? How does it perform when facing an adversarial attack?

AI RMF: Utilization of "clear and plain language that is understandable by a broad audience" as well as an emphasis on the need for communication of AI risks across an organization, between organizations, with customers, and to the public at large. (Pg 4)

- ***CalypsoAI:*** As AI becomes democratized, stakeholders must both understand their model's performance and be included in the TEVV process, which will enhance their understanding of AI risks throughout the AI/ML lifecycle and enable better organizational decision-making. It is also important to ensure that while stakeholders are involved in the risk management process, they must not slow it down. CalypsoAI addresses the need for communication of AI risk across an organization by providing a TEVV reporting mechanism with clear, easy-to-understand language, images, and graphs, enabling understanding and conversations about model risk between technical and non-technical stakeholders.

AI RMF: Acknowledgement that there is no one-size-fits-all approach to AI risk management. (Pg 4)

- ***CalypsoAI:*** Every time a model is deployed, the mission and the operational environment form a unique scenario populated with countless variables that the model could encounter and must be able to effectively address. As referenced above, context is key and a standardized, repeatable TEVV framework aligned to NIST's guidelines is critical to mitigate AI risk.

AI RMF: Acknowledgement that third-party data or systems present great challenges and that "the metrics or methodologies used by the organization developing the AI system may not align (or may not be transparent or documented) with the metrics or methodologies used by the organization deploying or operating the system." (Pg 8)

- ***CalypsoAI:*** Organizations frequently purchase pre-configured models, which means users only have the vendor's word that the model will perform as intended. Without knowing how the model is trained, explainability challenges arise, increasing the likelihood of the unintended consequences this framework seeks to address, such as model vulnerability to adversarial attacks and inaccurate performance in real-world conditions. Likewise, models developed in-house can lack consistent and easily-understood performance metrics across the model's lifecycle, which are necessary for confident deployment of AI models into any mission environment. Both scenarios pose significant risk to the organization's efficiency and effectiveness, and may cause undue harm to individual well-being. As suggested below, to mitigate these challenges, it is important that TEVV is conducted by an objective, independent third-party.

AI RMF: Acknowledgement of privacy and security concerns within AI systems. (Pg 10)

- ***CalypsoAI:*** In today's era of strategic competition, in which nation-states such as China and Russia use technology for authoritarian purposes, real-world AI deployment scenarios face privacy and security concerns. The model TEVV process must account for this. In addition to testing models against weather conditions, CalypsoAI's TEVV platform stress tests models against both privacy and adversarial attacks, giving stakeholders insight into how secure their models are.

AI RMF: Acknowledgement that "human judgment must be employed when deciding on the specific metrics related to AI trustworthy characteristics and the precise threshold values for their related metrics." (Pg 11)

- ***CalypsoAI:*** There is a place for subject matter experts (SMEs) and other stakeholders in the TEVV process because they understand the use case and mission conditions. As a result, SMEs should be the ones to use the independent TEVV findings to make an informed decision about whether or not to deploy a model. CalypsoAI's TEVV platform incorporates human judgment by allowing SMEs and mission-owners to align test parameters to deployment conditions and mission requirements at the beginning of a TEVV project.

Further suggestions:

- The NIST AI RMF should be more than a voluntary framework. While it is agreed that AI policy discussions are live and evolving, regulatory guidance from NIST would be more powerful in promoting trustworthy and responsible AI development and use in the United States. Requiring accountability in practices now, in the nascent stage of AI, will reap greater benefits in the future in terms of system safety, security, and transparency.
- The AI RMF states the need to identify "constraints in real-world applications that may lead to negative impacts." CalypsoAI suggests NIST elaborate on 1) how to test against real-world constraints, 2) how real-world constraints impact model performance, and 3) how to measure these impacts.
- CalypsoAI suggests highlighting the need to inform the AI/ML acquisition professional on how to approach the mission definition and intent with a focus on operational AI. The following must be defined:
    - Mission Improvement
    - Mission Operational Environment
    - Mission Function
    - Mission Temporal Requirement
    - Mission Failure Modes
    - Mission Monitoring and Enhancement

Conclusion:

CalypsoAI firmly supports NIST's effort in establishing a Risk Management Framework for Responsible AI and appreciates the opportunity to provide our thoughts and feedback on the path forward. We welcome any opportunity to work with NIST, industry partners, and broader

government agencies to assist in developing a responsible, trustworthy, and secure AI RMF for the benefit of all sectors.

For further questions or for more information, please do not hesitate to reach out to Hannah Mezei at