CALYPSOAI

CalypsoAI Response: NIST AI Risk Management Framework

Recognizing today's era of strategic competition, whereby nation-states such as China and Russia use technology for authoritarian purposes, it is imperative that U.S. artificial intelligence (AI) development and use reflects Western democratic values. As such, 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.

The AI RMF contains the necessary elements to manage risk with flexibility and can serve as an enduring resource. However, one piece that cannot be overemphasized is the need for rigorous testing, evaluation, verification, and validation (TEVV). This process is key to building trust into AI models, which will ultimately enable widespread AI adoption. It is also the best protection the U.S. has to mitigate risks associated with AI, such as resilience, explainability, and privacy. Hence, our comments all amplify the need to incorporate TEVV more frequently throughout the AI/machine learning (ML) lifecycle, particularly prior to deployment.

CalypsoAI agrees that standards will continue to evolve with the technology landscape. However, this should not require the creation of a cumbersome validation process that requires sign-off from multiple stakeholders each time we seek to deploy AI models, or cause a delay in creating a standardized validation method. Given CalypsoAI's expertise in third-party AI/ML model validation, we know that it is possible to institutionalize an automated TEVV process that mitigates risk and builds trust.

Institutionalizing a standardized process for independent AI/ML TEVV *prior to* – as opposed to only after – model deployment can address this issue. Oftentimes, organizations purchase algorithms that are already pre-configured, meaning users only have the vendor's word that it will perform as intended. Without knowing how the model is trained, explainability challenges arise, which heightens 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, algorithms that are developed in-house 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 a significant risk to the organization's efficiency and effectiveness, and may cause undue harm to individual well-being.

Since independent AI TEVV is currently missing from the NIST AI RMF, CalypsoAI has identified areas where the draft language can be updated to ensure this crucial step in the AI/ML lifecycle is not overlooked:

AI RMF: "Operators and evaluators provide monitoring and formal/informal test, evaluation, validation, and verification (TEVV) of system performance, relative to both technical and socio-technical requirements. These stakeholders, which include organizations which operate or employ AI systems, use the output for decisions or to evaluate their performance. This group can include users who interpret or incorporate the output of AI systems in settings with a high potential for adverse impacts. They might include academic, public, and private sector researchers; professional evaluators and auditors; system operators; and expert end users." (pg. 4)

- ***CalypsoAI:*** As AI becomes democratized, it is increasingly important for all stakeholders to both understand their model's performance and be included in the TEVV process. However, while continuous monitoring is vital to a model's success, it is equally if not more important for a model to be rigorously tested and validated *before* it is deployed. If AI/ML models are not tested during the procurement process, it is possible that vulnerabilities or inaccuracies in the models may go undetected. Consequently, we recommend updating this category to include model developers, who should perform a separate TEVV process so that they can confidently advance robust models to the operators and evaluators. This will also enhance understanding of AI risks throughout the AI/ML lifecycle, which will enable better organizational decision-making. It is also important to ensure that while all of these stakeholders are involved in the risk management process, they also should not slow it down.

AI RMF: "Validity of AI, especially machine learning (ML) models, can be assessed using technical characteristics. Validity for deployed AI systems is often assessed with ongoing audits or monitoring that confirm that a system behaves as intended. It may be possible to utilize and automate explicit measures based on variations of standard statistical or ML techniques and specify thresholds in requirements. Data generated from experiments that are designed to evaluate system performance also fall into this category and might include tests of causal hypotheses and assessments of robustness to adversarial attack." (pg. 8)

- ***CalypsoAI:*** Since AI is meant to enhance human performance, finding ways to automate tasks is key to harnessing the technology's potential. As it currently stands, the testing and evaluation (T&E) process is labor intensive. As such, automating this process both pre- and post-deployment gives data scientists valuable time back and shifts dependence away from arbitrary model evaluation metrics, such as F1 scores, ROC, AUC, Precision, and Recall. This is important because these metrics only offer insight into how a model performs on its training data. In order to be effective, TEVV must include metrics that account for model performance on unseen data, which will help determine the model's real-world performance before it is deployed. For example, if TEVV only performs on its training data, an algorithm that is used in a UAV over Afghanistan will not perform the same over Ukraine because the ground conditions vary. The model must account for these differences.

Given this section addresses technical characteristics – which the AI RMF notes are "factors that are under the direct control of AI system designers and developers," – it is a good opportunity to highlight the importance of pre-deployment validation. Moreover, this section should focus on developing guidance for organizations to determine acceptable model thresholds for their specific conditions and risk factors. This will enable them to choose tests that are automated, adaptable, and scalable which aligns with this AI RMF statement: "Determining a threshold for accuracy that corresponds with acceptable risk is fundamental to AI risk management and highly context-dependent."

AI RMF: "Figure 6: Risk management should be performed throughout the AI system life cycle to ensure it is continuous and timely. Example activities for each stage of the AI lifecycle follow. Pre-Design: data collection, curation or selection, problem formulation, and identification of stakeholders. Design & Development: data analysis, data cleaning, model training, and requirement analysis. Test & Evaluation: technical validation and verification. Deployment: user feedback and override, post deployment monitoring, and decommissioning." (pg. 15)

- ***CalypsoAI:*** Since "risk management should be performed throughout the AI system life cycle to ensure it is continuous and timely," testing and evaluation should not only be a step in the process; rather, it should be ongoing and repeatable. The "Test & Evaluation" category makes this ambiguous. Consequently, we recommend changing the category to "Independent Test & Evaluation," which creates a clear step in the risk management process that is necessary for safe deployment.

AI RMF: Table 2, Example of categories and subcategories for Measure function: "**2. Systems are evaluated.** Accuracy, reliability, robustness, resilience (or ML security), explainability and interpretability, privacy, safety, bias, and other system performance or assurance criteria are measured, qualitatively or quantitatively. Mechanisms for tracking identified risks over time are in place, particularly if potential risks are difficult to assess using currently available measurement techniques, or are not yet available. **3. Feedback from appropriate experts and**

stakeholders is gathered and assessed. Subject matter experts assist in measuring and validating whether the system is performing consistently with their intended use and as expected in the specific deployment setting. Measurable performance improvements (e.g., participatory methods) based on consultations are identified." (pg. 17)

- ***CalypsoAI:*** If the testing and validation process is designed such that it is standardized, there should be no need for a separate step that includes subject matter expert feedback. While conversations about measurable performance improvements should be ongoing at a societal level, adjustments need to be made real-time for specific conditions. Otherwise, this may hinder AI adoption or use. At the same time, there is a place for subject matter experts and other stakeholders because they understand the use case and mission conditions, such as the amount of fog to expect or risk of vulnerability to adversarial attacks. As a result, they should be the ones to use the Independent T&E findings to make an informed decision about whether or not to deploy an algorithm.
  Additionally, automating the process removes subjectivity on a case-by-case basis while saving time and money. Knowing that red teaming for AI/ML TEVV typically requires a minimum of 8-15 months and is conducted in an ad hoc fashion, an automated TEVV process significantly reduces costs and the time to perform TEVV from months to days or even hours.

AI RMF: "Management can take the form of deploying the system as is if the risks are deemed tolerable; deploying the system in production environments subject to increased testing or other controls; or decommissioning the system entirely if the risks are deemed too significant and cannot be sufficiently addressed. Like other risk management efforts, AI risk management must be ongoing." (pg. 17)

- ***CalypsoAI:*** The current AI deployment gap heightens the importance of this point. According to Gartner, "85 percent of AI projects will deliver erroneous outcomes due to bias in data, algorithms or the teams responsible for managing them" by the end of this year. In order to build trust in AI systems, safe deployment is essential. This requires users to both rigorously test and validate their models before deployment into production, as well as continuing to validate models once they are deployed. This will greatly reduce any "risks" that may arise when determining whether to deploy these systems.