CALYPSO AI

CalypsoAI's Response to the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework Request for Information

August 19, 2021

CalypsoAI is pleased to respond to the National Institute of Standards and Technology (NIST)'s Artificial Intelligence (AI) Risk Management Framework (RMF) request for information (RFI). We believe an AI framework created by NIST and informed by academia, industry, and government, is an important step towards achieving the goal of improving the management of risks to individuals, organizations, and society associated with AI.

As a nascent capability, AI/ML products are being developed with minimal standard tooling, leading to algorithms of uncertain quality, subjective trustworthiness, and potential vulnerability to attack. Organizations across all sectors lack the tools to adequately assess and monitor AI/ML products, which often means these solutions are not deployed. CalypsoAI addresses these concerns.

As a software company and thought leader for Secure and Trusted AI, CalypsoAI knows firsthand the importance of an AI Risk Management Framework. Our work focusing on the testing, evaluation, verification and validation (TEVV) of AI and Machine Learning (ML) models has resulted in a contract with a Department of Homeland Security (DHS) Science and Technology (S&T) Office's Screening at Speed Program (SaS) with work tangential to NIST entitled Software for Trusted Intelligence. It has also earned us a contract with the Secretary of the Air Force Concepts, Development, and Management Office (SAF/CDM) entitled Secure Artificial Intelligence.

Drawing from our cutting-edge AI research and development (R&D) team and experience working with industry and government in the AI space, CalypsoAI provides the following comments to inform NIST's consideration of options and development of an AI RMF.

1. Response to Specific Requests for Feedback

1.1 Greatest Challenges in Improving How AI Actors Manage AI-Related Risks

There are five key challenges to improving how AI actors manage AI-related risks:

1. a lack of common understanding between AI actors,
2. siloed AI/ML initiatives,
3. insufficient regulation and oversight,

4. scarcity of skilled talent, and
5. the pace of which AI research far surpasses the pace of implementation

In this line of work, it is important to recognize that the unfolding pace of research is faster than the pace of actualized production across all sectors. Additionally, the experimental modes of implementation are at odds with highly structured risk management systems. All of these factors necessitate a more adaptable approach within specific bounds, which we outline below.

First, without a common understanding between actors, AI-related risk cannot be properly managed. AI jargon is often not understood by key decision-makers, who assess the real impact of the risks associated with AI and are responsible for determining the level of acceptable risk. Furthermore, AI teams and vendors typically work on a discreet part of an overall operational problem. They target AI-specific metrics, such as high performance or accuracy, without necessarily understanding the complete context in which models will run.

Second, within the U.S. Government (USG), AI/ML initiatives continue to be siloed, with isolated actors attempting to build and deploy models with limited access to standard industry tooling. This often results in inefficiencies and redundancy in AI efforts, as well as a lack of communication across USG organizations on best practices.

Addressing these siloes will require a greater investment of resources. However, as resource investment increases, greater regulation and oversight is needed. Without regulation and oversight, commercial AI actors typically establish a working solution first and deal with issues later as they arise. In contrast, government organizations often look for a clear path or standard procedures before acting. While this is good practice, it may also lead to slowed progress or inaction. Consequently, oversight in the form of auditability and regulations for risk management, coupled with incentivization, will ensure AI actors dedicate the resources necessary to appropriately manage AI-related risks. Sound oversight includes established evaluations, measures, and considerations for each of the different areas of interest: accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security, and mitigation of unintended and/or harmful bias. These all contribute to ensuring accountability and AI model robustness which, although more than a checklist, creates parameters for established good practices.

At the same time, we must grow the pool of good-faith Data Science experts, who possess the foundational knowledge to properly address AI risk management. Without this talent, we will face additional challenges with implementing an AI risk management framework.

Finally, the pace of research far surpasses the pace of implementation. Most companies that are attempting to integrate AI/ML into their day-to-day operations are not successful because most executives do not truly understand ML. For example, it is common – in all sectors – for governments to communicate their needs through scenarios and case studies, with little to no expertise on how to translate broad scenarios into the granular scope of a ML problem. Hence, among the technical professionals, there is a need for greater education on the problems, actors, and systemic consequences of their work beyond a simple end-user experience. While human-centered design research and end-user engagement is important within the scoping process, successful risk management and AI/ML delivery requires an understanding of the human systems in which the models are to operate. At present, this demands the technical support staff to have on-hand product managers, product researchers, and product designers to collaborate with data scientists and the federal problem-owner.

Today, the ability to create a machine learning model is trivial and the ability to have model confidence is moderately accessible. Yet the ability to design a service, transaction, or system of technologies that integrate with ML for problem-solving is rare. Efforts to safeguard AI/ML adoption are difficult to design and implement in light of this broader set of conditions, which is why developing this AI Risk Management will be instrumental to our success.

1.2 How Organizations Currently Define and Manage Characteristics of AI Trustworthiness – Important Characteristics

One way CalypsoAI defines and manages characteristics of AI trustworthiness, among other characteristics, is through the insights gleaned from our support to customers, particularly when testing for model robustness by evaluating model performance on data samples with added synthetic noise. In our internal research, we have identified that highly performant models demonstrate inconsistent degradations when introduced with different types of noise, whereby inference on samples with one type of noise cause almost no performance loss in one model, but the performance of a second model is reduced to that of random guessing.

We partially define trustworthiness by assessing a model’s explainability and interpretability, for which there are many different means of gaining insight into model performance. A few examples include adding an attention layer for recurrent neural networks, applying a saliency map for convolutional neural networks, and the more general method for local interpretable model-agnostic explanations (LIME). However, the utility of these methods can vary widely and are often subjective. Consequently, they will have limited utility for expressing model interpretability to a broad audience, to include senior executives and those who are not AI professionals. Despite these limitations, these methods for gaining insights into models can serve as a valuable point of model interrogation that can be utilized for AI audits and governance.

While it is generally known that simpler models built on fewer features lead to more secure and interpretable models, there is a strong allure that a highly complex, black-box model will better capture hidden relations in the data. Therefore, in order to promote AI trustworthiness, the RMF application should focus on simple models that highlight the risks of more complicated ones. Although not a perfect measure, one means of quantifying model complexity is using the number of tunable parameters in the model.

1.3 How Organizations Currently Define and Manage Principles of AI Trustworthiness – Important Principles

As a matter of principle, quantifiable measurements of AI/ML performance and risk that can be clearly evaluated and/or tracked should take precedence over assuming AI developers perform necessary due diligence. During a panel titled, “How does explainable AI fit into the trustworthy AI ecosystem?” at the NIST Explainable AI Workshop, a virtual event held January 26-28, 2021, CalypsoAI asked the panelists (from Google, Bank of America, and Microsoft) what, if any, metrics they use to establish a model as explainable and trustworthy. The panelist's response was that “trust is earned” and demonstrated through “customer adoption and satisfaction.” This expresses an idea that trustworthy AI is really a matter of trusting the team or individual who develops the model that they are working in good faith to apply due diligence and avoid potential pitfalls. While it is certainly helpful that consumers and regulators trust the individual, team, or organization that is developing and deploying an AI model, this developed trust is an intangible measure that cannot be clearly evaluated or tracked.

Another important principle for the implementation of trustworthy AI is to consider the full model lifecycle over time, and the conditions of human machine interaction that foster model trust. For example, research into autonomous vehicles demonstrates the importance of conversational interactions between the technology and the user. The combination of visual, gestural, tactile, and verbal interfaces to engage and capture a response from the system are important. If the user articulates “drive to the Capitol,” we should expect an algorithm to respond with a verbal “Washington DC?” a touchscreen to show the map, and possibly sensor cues such as highlighting the destination on the map. The interaction should require engagement for confirmation, such as “please say Washington DC to confirm your destination,” guiding the user into dialogue.

This example highlights that trust in AI is dependent upon an ecosystem of multi-modal interactions. Trust is not an “invisible glue” or a leap of faith, it is a cognitive construction built over time through sensory engagement. At present, we see few organizations fully adopting these opportunities for trustworthy AI/ML, and the majority of insights remain in academic research, sometimes under the banner of second-order cybernetics, pioneered by human computer interaction leaders such as Jodi Forlizzi, John Zimmerman, and Paul Pangero.

1.4 How to Identify, Assess, Prioritize, Mitigate, or Communicate AI Risk

Today, many organizations are blind to AI/ML risk. They trust the data, the data scientist, and the model results by default. Even leadership with USG agencies have been heard to say, “the models perform better in the field than in the test bed,” or, “models are already accurate.” Bias is widely recognized as a model risk, yet beyond that, there is little awareness of problems such as data shift or model drift until a problem has taken place.

While not perfect, there are several established best practices and methodologies that exist for creating high quality ML models. Examples include applying cross-validation, carefully considering bias and variance, removing co-correlated features, requiring review processes, and versioning data sets, models, and environments. An RMF application should capture and report the best practices that are incorporated into the model development process.

1.5 What NIST Should Consider to Ensure that the AI RMF Aligns with and Supports Other Efforts

CalypsoAI is committed to developing frameworks that support developing and deploying machine learning models with maximum accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security, and mitigation of unintended and/or harmful bias. Our work toward this goal is made available in the following two applications:

- VESPR: a full secure machine learning development lifecycle application that makes machine learning verification and validation testing an integral part of the development process. This application support developing classification models from either tabular or image datasets, and regression models from tabular datasets.

- Fortika: a test suite for evaluating image classification and object detection algorithms beyond the standard correctness metrics of baseline performance. This application is focused on leadership and senior executives, bridging the technical gap between machine learning engineers and decision makers.

1.6 AI Benefits and Issues Related to Inclusiveness

Commercial organizations cater to groups in a way that they believe will maximize revenue. Therefore, for-profit companies frequently employing exclusive practices, even unintentionally. This includes both making sure their products and services meet the needs of the most lucrative groups, and avoiding a negative image caused by neglecting other populations. To the end, one of the best mechanisms for promoting inclusiveness in AI design is by making organizations accountable through regulation, reporting, and oversight.

Second, collaborative and participatory processes commonly used within civil society are equally applicable to model development. The creation of data, the curation of a feature set, model weights, and model training can be collaborative. Reviews on model performance relative to test results and model trade off analysis to down select a model for implementation do not have to be an expert-led practice. Dialogue mapping, in particular, can lend itself toward participatory model design. For collaborative AI/ML to succeed, there is a demand for collaborative design expertise, and often a need to get the data science questions and decisions away from the technologies. A collaborative session with a python notebook will not succeed, whereas a sequence of smaller workshops around data collection, model factors and weighting, and model selection, can open the forum while providing ample opportunity for technical experts.

At CalypsoAI, we utilize this approach within the proven product management and delivery system of Discovery, Framing, Alpha, Beta, Live. Working with our stakeholders, we undertake participatory decisions and rapid prototyping exercises to quickly bring together diverse stakeholders, advance and share insights, and optimize our ML solutions.

1.7 Appropriateness of the Attributes NIST has Developed for the AI RMF

Attribute 3, “Use plain language that is understandable by a broad audience, including senior executives and those who are not AI professionals, while still of sufficient technical depth to be useful to practitioners across many domains,” touches on a critically important aspect of an RMF that will foster widespread adoption. In order to provide “sufficient technical depth” to be actionable in a consistent and meaningful way, the RMF will also need to include concrete metrics with mathematical backing. This highlights a need for a balance, as many technical details may not be expressed in sufficiently plain language for a broad audience to understand the details.

Attribute 5 mentions that the RMF be “voluntary, and non-prescriptive”. While there is a difference between an RMF and regulations, the RMF will need to address and facilitate compliance with regulatory requirements to gain widespread acceptance.

1.8 Effective Ways to Structure the RMF

NIST should look to the USG 18F Documentation on agile product development within the USG as a coherent framework that can be rapidly adopted by stakeholders. The composition of artifacts to facilitate framework implementation, the ability to connect multiple stakeholders, and the realization of concrete outcomes demonstrates the 18F as a model for replication.

Complimentary to that framework, NIST should recognize that a consistent failing for the successful implementation of AI/ML within the USG is the disconnect between government scale problems, the subject matter expertise of the staff, and the continued necessity for processes such as human centered design within AI/ML scoping.

Notably, human centered design is insufficient for the development of successful AI/ML tools. It is essential that program managers make available opportunities for exploratory data analysis and experiments with low-risk proxy data, while simultaneously connecting vendors with the field deployment agents who will be impacted by the model performance and reliability.

These steps are essential because whereas the scoping and deployment of ML requires abstract, deductive reasoning that is often out of sync with available expertise, supporting data scientists often maintain an overly narrow focus on model performance, data quality, and insular model testing void of actual real-world conditions or on-site expertise.

1.9 The RMF Advancing a Skilled Workforce

An AI RMF will be instrumental in advancing a skilled workforce for the following reasons: By providing AI guardrails, NIST is improving overall organizational awareness of conditions that may cause models to fail without depending on extreme domain expertise.

To successfully realize AI/ML initiatives with a given program management office (PMO), vendor participation needs to extend beyond simple commercial off-the-shelf (COTS) acquisitions or delivery of isolated models. It is fundamental that the vendor is a recognized stakeholder, and vendors that hold both technical and subject matter expertise are critical to scope, design, and implement successful programs.

However, there are limits to the reliance on contractors and vendors. While their AI/ML expertise is valuable, their contributions also drive organizational failures in this space. PMOs remain constrained in their ability to systematically build successful ML-driven initiatives due to the over-reliance on the contractor community. At present, all insights, best practices, and value derived from a given AI/ML initiative remains with the vendor, meaning it also disappears with the vendor. To satisfy the demand for vendors, yet further drive the greatest levels of organizational return on investment, it is essential that AI capabilities are productized for federal use. Products can be delivered by USG vendors or developed for intentional creation as federal intellectual property (IP).

CalypsoAI's product delivery thus far suggests that vendor IP ownership that comes with the freedom of "pay as you go" product offerings provide USG PMOs optimum flexibility, scalability, and impact.

The utility of the Framework for recruitment, hiring, and development will depend on the level of adoption. This will be largely driven by how well it supports practitioners in better providing value and organizations in complying with regulations.

For practitioners, adoption will depend on how much the Framework serves as a source of training or a reference source, or by opening doors with potential employers. This is similar to cybersecurity certifications like the NIST Cybersecurity Framework (NCSF), CompTIA Security+, CISSP, CISM, etc.

As discussed previously, incentivization and accountability through reporting, regulations, and auditing will motivate organizations to allocate the resources needed to enact change.

1.10 Governance Issues

To advance government adoption, USG Agencies should not -- and must not -- create technical review boards (TRBs) for agency-level decisions on project development, design, implementation, and manufacturing and engineering. This antiquated approach has proven itself too slow and ineffective for the fast-paced nature of emerging technologies, and the complex environments of government interests. Furthermore, the membership of TRBs is traditionally composed of subject matter experts with traditional perceptions of project management, who impose waterfall methodologies, a reliance on Ghant charts, and similar tools at odds with modern software development.

To advance governance, USG agencies are better positioned to adopt broad frameworks which can be modified to suit local needs and goals. Within a given agency, there is a demand for a team of individuals with expertise in modern product development, ML, and software development to work together and steer internal efforts, acquisitions, and bring continuity to PMO implementations of AI/ML initiatives.

The NIST RMF document indicates that "Stakeholders include but are not limited to industry, civil society groups, academic institutions, federal agencies, state, local, territorial, tribal, and foreign governments, standards developing organizations and researchers." To appeal to this broad range of stakeholders, the Framework will need to support regulation at the local, national, and international level. The Framework should be developed in collaboration with regulatory agencies across those levels, so it can be positioned to enable practitioners and organizations to comply with those regulations.

2. Conclusion

CalypsoAI firmly supports NIST's effort in establishing a Risk Management Framework for Artificial Intelligence and appreciates the opportunity to provide our thoughts and feedback on the path forward. We welcome any opportunity to work with NIST, industry, 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 Dr. Mitchell Situka-Sipus at Mitchell.Sipus@CAILab.com.