BLUVECTOR®
A COMCAST COMPANY

BluVector Comments on NIST's RFI on Artificial Intelligence Risk Management Framework

BluVector is an Arlington VA cybersecurity company that provides cybersecurity solutions that allow government and businesses to manage risk and operate with greater confidence that their data and systems are protected. BluVector has over a decade's experience innovating in the areas of machine learning and artificial intelligence. We develop and deploy AI solutions across global government and commercial networks to solve some of the most challenging cybersecurity problems. and our AI-powered technology recently won the U.S. Cyber Command Competition (link to press release).

BluVector offers these comments in response NIST's request for information on an *Artificial Intelligence Risk Management Framework*

Response to RFI Question #1 The greatest challenges in improving how AI actors manage AI-related risks – where "manage" means identify, assess, prioritize, respond to, or communicate those risks;

The application of AI to cybersecurity highlights unique, important and difficult challenges that should be incorporated into a framework for managing risks. Cybersecurity applications require rapid inspection and analysis of incredibly large volumes of heterogeneous data generated by both humans and machines that evolve in time and can significantly vary for each deployment.

AI allows us to sift through these large volumes of data with unprecedented efficiency and effectiveness and but can also pose new risks if improperly developed or naively deployed. Furthermore, AI applications must also contend with maintaining effectiveness in environments where adversaries are actively seeking to defeat or avoid detection and can be equally aware of the potential pitfalls or risks of the AI systems they are trying to defeat.

Framework recommendation: The greatest or significant challenges that need to be incorporated into the AI RMF should include:

a. Accuracy against the unknown
b. Operation across radically different and/or changing environments.
c. Risk against direct attack.

Accuracy against the unknown: AI models are trained on items that they are meant to identify when deployed. Testing the accuracy of these systems is critical, and in most cases, generally straightforward. A significant challenge for the Framework are AI models that are relied upon to classify brand new instances of items, based on shared characteristic patterns, that have not been previously seen or identified before and therefore cannot be represented in the training or test data.

An example of this is the detection of zero-day malware intrusion attempts. Most malware detection software relies largely on signature-based methods of known malicious code that can cause harm to computer systems, but zero-day malware is, by definition, unknown and unidentified as malware. Consequently, AI classification models must identify unknown, zero-day malware while maintaining high accuracy, reducing both false negatives and false positives. A framework for managing the accuracy and reliability of such an AI model, against the unknown, requires additional consideration of test sets and is an important challenge to consider. (see: US Patent 9,665,713, "*System and Method for Automated Machine-Learning, Zero-Day, Malware Detection*", Assignee BluVector. https://patents.google.com/patent/US9665713B2)

Risk against direct attack. Another difficult challenge for the Framework is the risk of operation of an AI system in an environment where there are likely to be adversaries trying to defeat, circumvent or even alter an AI system. This is an additional risk that is highly prevalent in cybersecurity but also important in many other industries, such as financial, law enforcement, industrial control etc. An AI system that is accurate and reliable but can be reverse engineered or cheated presents a new risk.

One of the greatest challenges for a successful Framework is balancing competing risk factors. For example, "explainability", an important characteristic of AI trustworthiness, must be balanced with protecting the model itself. Additionally, BluVector has found that shared AI model across all customers is more susceptible to cyber adversarial attacks. Tailoring models to each customer provides a moving target defense and while improving efficacy it also creates a more challenging defense for cyber adversaries to counter.

Response to RFI Question #4: The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management – including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

For all cybersecurity applications including those that use AI, the risks are intimately related to an organization's overarching enterprise risk management. Even in cases where cybersecurity practices are mandated by regulation, each organization uniquely determines how it will secure the operation of its network, protect digital assets, and maintain data/user privacy. Ultimately the choices of resources, tools and practices employed are driven by the organization's overall tolerance of risk. The desire to mitigate risk by using an AI-enabled cybersecurity application(s) will have a direct impact on overarching risk of the enterprise and must be managed accordingly.

Framework recommendation: Organizations must be able to manage characteristics of their AI models to align with their operational needs and tolerance for risk. BluVector recommends that the Framework address user control of model parameters for risk management and provides guidance for users who will utilize these adjustments to better match their organizational objectives. Framework guidance can include sharing specifications on relationship between parameters and model performance along with safeguards to prevent unintended behavior.

For example, the Receiver-Operating Characteristics of a machine-learned binary classifier will determine the sensitivity of that classifier. More sensitivity comes at the cost of more false positives and visa versa. Furthermore, differences in data distributions across different implementations might result in different classifier characteristics. BluVector has found that users of our Zero-Day intrusion detection classifiers benefit from the ability to adjust model tolerances in a manner consistent with their security policies and analyst workflows. These adjustment are evaluated regularly over time as network behavior changes or conditions around organization risk changes.

Response to RFI Question #5: Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

The lifecycle of a supervised machine learning model consists of training, testing, followed by implementation. The training-testing phase of the cycle will typically be conducted remotely by the developers of the AI model. Many current approaches to machine learning use algorithms that require a static and complete training data set where all representative data samples are available during training. BluVector finds that operational data seen in the field may have meaningful differences from the static training data.

Existing machine learning techniques disclose learning algorithms and processes but do not cover methods for augmenting or retraining classifiers based on data not accessible to the original trainer. Furthermore, machine learning models do not enable training on data samples that an end user does not wish to disclose, due to proprietary or privacy concerns, to a 3rd-party which was originally responsible for conducting the machine learning.

BluVector provides its non-expert users a method for batched, supervised, in-situ machine learning classifier retraining for malware identification and model heterogeneity.The method produces a fully trained parent classifier model to organizations, operating in different locations, with the capability for in-situ retraining of the system by augmenting the training data with the organization's proprietary or private data. (see: US Patent 10,121,108; "*System and Method for In-Situ Classifier Retraining for Malware Identification and Model Heterogeneity*", Assignee BluVector Inc. https://patents.google.com/patent/US10121108)

Framework recommendation: The Framework should incorporate the ability for non-expert end users to augment a pre-trained model with additional data including the validation and testing of any updated model against risk criteria before it is implemented.

BluVector offers its method for in-situ training as a guiding example for a more general framework. Our in-situ training method:

1. adjudicates the class determination of the parent classifier over the plurality of the samples evaluated by the in-situ retraining system or systems,
2. determines a minimum number of adjudicated samples required to initiate the in-situ retraining process,
3. creates a new training and test set using samples from one or more in-situ systems,
4. blends a feature vector representation of the in-situ training and test sets with a feature vector representation of the parent training and test sets,
5. conducts machine learning over the blended training set.
6. evaluates the new and parent models using the blended test set and additional unlabeled samples, and elects whether to replace the parent classifier with the retrained version.