AEROSPACE

REGIONAL OFFICES:
14745 Lee Road
Chantilly, VA 20151

The Aerospace Corporation
CORPORATE OFFICES:
2310 E. El Segundo Blvd.
El Segundo, CA 90245-4609

www.aerospace.org

Transmitted via Electronic Mail

19 August 2021

National Institute of Standards and Technology
100 Bureau Dr, Gaithersburg, MD 20899

Attention: NIST AI Risk Management Framework (AIframework@nist.gov)

Subject: The Aerospace Corporation response to NIST RFI "Artificial Intelligence Risk Framework"

Reference: Request for Information (RFI) – Artificial Intelligence Risk Framework

Dear Ms. Tabassi,

The Aerospace Corporation (Aerospace) wants to thank you for the opportunity to respond to your Request For Information (RFIs) on Artificial Intelligence (AI) Risk Management Framework (RMF) (due 8/19/21).

As you know from our earlier interaction, dialogue, and input on the NIST effort on AI Standards, directed by an earlier Presidential Executive Order on AI, Aerospace is a Federally Funded Research and Development Center (FFRDC) with a 60+ year heritage providing Mission Assurance for the Space Enterprise. With the increasing adoption and interest in AI within the space community, extending our Mission Assurance offerings to AI-enabled space systems is a natural evolution.

In response to the AI RMFRFI, Aerospace proffers two high quality resources:

A. "A Framework for Trusted Artificial Intelligence in High-Consequence Environments": This document presents a solid framework for Trusted AI in use cases where there is high risk and little margin for error, such as space and related complex systems, a topic area where we are the premier FFRDC for the U.S. Government and their private sector partners. This should be a truly informative piece for developing the AI Risk Management Framework (RMF) agnostic and regardless of domain.

B. "A Framework for Developing Trust in Artificial Intelligence": This is a publicly released issuance that describes the relevance of Aerospace's AI RMF framework model to policymakers.

AEROSPACE

REGIONAL OFFICES:
14745 Lee Road
Chantilly, VA 20151

The Aerospace Corporation
CORPORATE OFFICES:
2310 E. El Segundo Blvd.
El Segundo, CA 90245-4609

www.aerospace.org

For further consideration in future engagements, we have another document, currently in final stages of development, that is pertinent to this RFI which may be made available upon public release.

C. "Trusted Artificial Intelligence for Flight Infusion and Science Service": A cooperative partnership between Aerospace and JPL to: (1) apply the Aerospace framework to two JPL space exploration programs and collect lessons learned, (2) develop best practices for future JPL AI/ML autonomy software to increase trust.

Thank you for this opportunity to collaborate with you and your office as a trusted and valued partner. We also want to congratulate you on being appointed to the White House Task Force on AI Research Resource (we'll respond to that RFI separately). It is an honor to work with you and we compliment you on your appointment. We stand by to further collaborate and partner with you on future AI-related efforts.

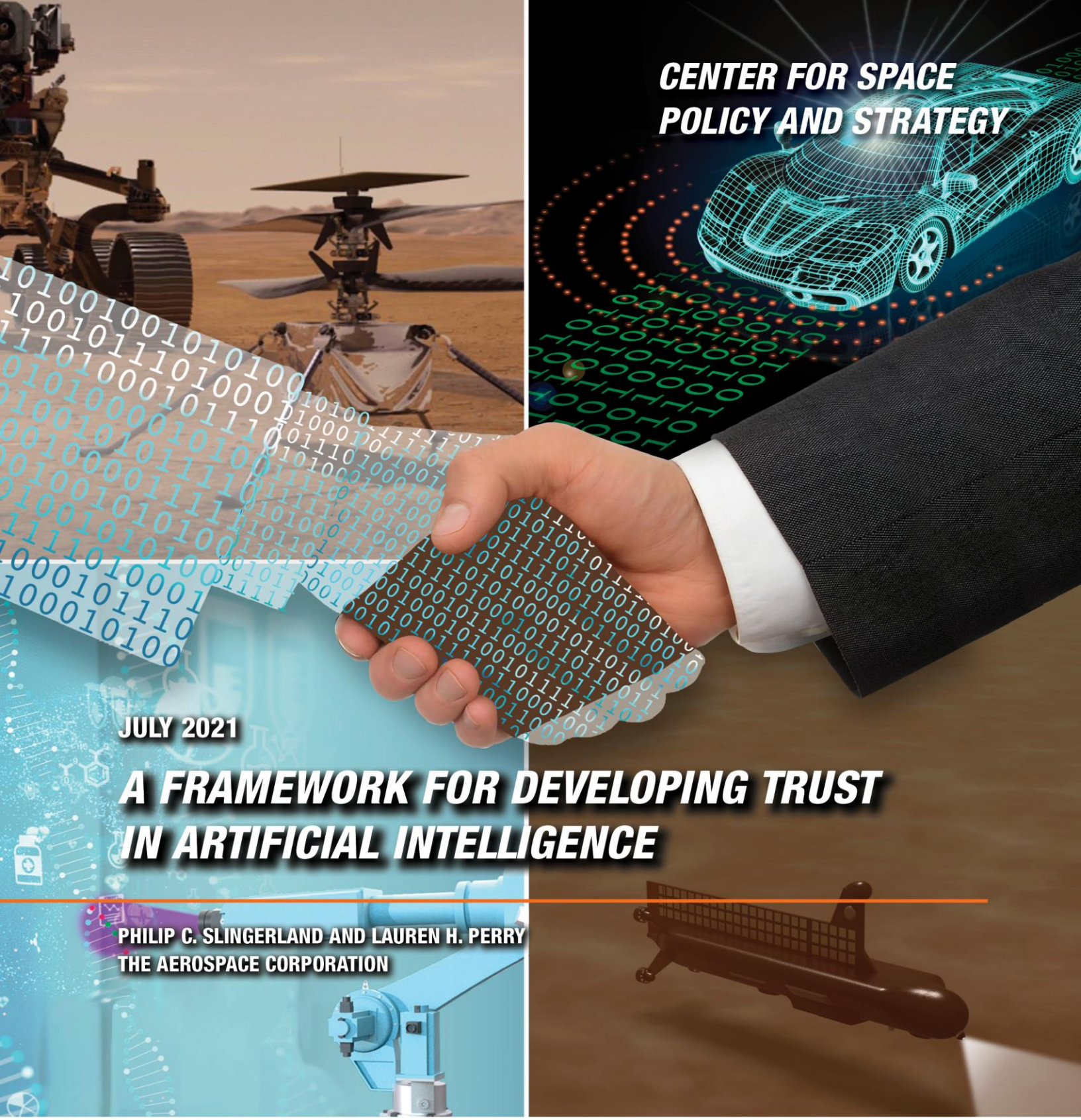
Any questions regarding this response or attachments can be directed to Mr. James M. Myers at 703-812-2012 or via email at james.m.myers@aero.org.

V/R

James M. Myers
Vice President, Civil Systems Group
The Aerospace Corporation

Attachments:

1. A Framework for Trusted Artificial Intelligence in High-Consequence Environments (Filename: "A Framework for Trusted Artificial Intelligence in High-Consequence Environments.pdf")
2. A Framework for Developing Trust in Artificial Intelligence (Filename: "A Framework for Developing Trust in Artificial Intelligence.pdf")



JULY 2021

A FRAMEWORK FOR DEVELOPING TRUST IN ARTIFICIAL INTELLIGENCE

PHILIP C. SLINGERLAND AND LAUREN H. PERRY
THE AEROSPACE CORPORATION

AEROSPACE

DR. PHILIP C. SLINGERLAND

Dr. Philip C. Slingerland is a senior engineering specialist in the Machine Intelligence and Exploitation Department at The Aerospace Corporation. Slingerland's work focuses on machine-learning and computer vision projects for a variety of intelligence community, DOD, and commercial customers. Previously, he spent four years as a data scientist and software developer at Metron Scientific Solutions in support of many Naval Sea Systems Command (NAVSEA) studies. Slingerland has a background in sensor modeling and characterization with a Ph.D. in physics, studying the performance of terahertz quantum cascade lasers (QCLs) for remote sensing applications.

LAUREN H. PERRY

Lauren H. Perry is a senior project engineer in the Space Applications Group at The Aerospace Corporation. Her work incorporates AI/ML technologies into traditional software development programs for the intelligence community, DOD, and commercial customers. Previously, she was the analytical lead for a DOD project established to improve joint interoperability within the Integrated Air and Missile Defense (IAMD) Family of Systems and enhance air warfare capability. Perry was also a reliability engineer at Lockheed Martin Space Systems Company. She has a background in experimental design, applied statistics, and statistical engineering for the aerospace domain.

ABOUT THE CENTER FOR SPACE POLICY AND STRATEGY

The Center for Space Policy and Strategy is dedicated to shaping the future by providing nonpartisan research and strategic analysis to decisionmakers. The center is part of The Aerospace Corporation, a nonprofit organization that advises the government on complex space enterprise and systems engineering problems.

Contact us at www.aerospace.org/policy or policy@aero.org

AEROSPACE

Summary

Artificial intelligence (AI) is a critical technology within a wide array of applications that is increasingly impacting people's lives. New AI-based capabilities have enhanced or enabled technologies that previously were not thought possible. However, with AI's increased presence has also come rising concern. This is especially true in domains where the degree of risk or the potential for harm is high. Additionally, as more people engage with AI in their personal and professional lives, human perceptions and trust of AI will increasingly influence how AI-based applications are deployed. As a result, there is a strong demand to understand both the opportunities and concerns regarding AI's use. Policymakers need tools at their disposal to assess how much investment in AI is needed to reach the right level of trust and resiliency, and what should be demanded of AI to build trust with users. This document collects a set of definitions and best practices into a framework that spans the lifecycle of AI development. By breaking down and highlighting the challenges of each AI development phase, policymakers can see what aspects of trusted AI relate to their domain and how to achieve their vision of an increasingly AI-enabled future.

Introduction

Artificial intelligence (AI) has become a critical technology and central topic of discussion due to its well-established success in a wide array of applications. Success, however, has flagged due to the difficulty in ensuring humans can intervene in AI algorithms and understand how they operate in complex systems. In response, there is growing demand for trust in AI to address both the expectations and concerns regarding its use. But this need for trust is complicated by a public discussion that criticizes the misalignment between user expectations and the true capabilities of modern AI systems.[1] These discussions are based on valid concerns but are also clouded by inconsistent terminology used to define trusted AI. To address the varied requirements of AI-based applications and the lack of clear terminology, this document puts forth definitions and a framework to consider trust. It is based on concepts that cut across AI domains, with the intent to help policymakers quantify the risks and rewards of AI.

Many questions arise when considering if AI has a role within a particular domain. For example, consider a constellation of proliferated low earth orbit (pLEO) satellites that requires some level of autonomy to operate. Due to the complexity of implementing control and management software for a large fleet of satellites, AI may be an attractive option to relieve the burden. However, the many questions related to whether or not to employ AI in such a situation are challenging to answer. Stakeholders will typically ask: *What does AI have*

to offer? Will it work as intended when deployed? Have I done enough testing? How much risk am I taking on? Am I adding too much complexity? Underlying all of these is the fundamental question of whether AI is appropriate for a given mission or need. This creates a distinction between cases: 1) There is a need that is not met by current capabilities and that can only be enabled by AI, 2) AI offers a potential enhancement to existing capabilities, or 3) AI has already been deployed within an operational system and trust must be established post-deployment. The framework assists policymakers in these cases by providing clear definitions of trust and tools to answer some of the questions above. Above all, the framework strives to reduce the uncertainty of knowing whether AI is appropriate.

AI Versus ML

Artificial intelligence (AI) is a discipline within computer science that attempts to accomplish tasks that a human is capable of, but with software. Machine learning (ML) is a subfield of AI that learns from data how to accomplish the tasks of AI.

Trusted AI

AI capability that can provide reasonable confidence that it has satisfied user-defined objectives in a proper and interpretable way over its lifetime.

Trusted AI is defined here as an *AI capability that can provide reasonable confidence that it has satisfied user-defined objectives in a proper and interpretable way over its lifetime*. An AI that can be relied upon to operate safely in a high-consequence environment and to do no harm must be designed for trust from the start. The trusted AI framework is a means to assist with this design challenge. It is comprised of a set of best practices that recommend ways to incorporate trust into every phase of the AI lifecycle. When the framework is combined with existing processes (e.g., the definition of requirements, the construction of test plans, or as part of a user engagement study), it can help to balance the capabilities provided by AI with the additional challenges of real-world deployment. These concepts share and apply similar lessons to those of SecDevOps[*2], where DevOps practices are enhanced with an increased focus on security. SecDevOps recognizes that security considerations impact the entire process of delivering software applications. The trusted AI framework recognizes that trust impacts the process of developing AI-based applications and should be incorporated into existing DevOps practices tailored for machine learning (e.g., MLOps[†]). Ultimately, the framework provides policymakers the means to understand the challenges and required mitigations whenever AI will be deployed in a setting that impacts users, the external environment, or other systems.

The Current Landscape

Academic, commercial, and government sectors have increasingly studied and reported on topics centered around trusted AI. Initially, these studies focused on adversarial and explainable AI[3] (see sidebar on page 3). Adversarial AI[4] emerged from academia but was quickly elevated as a point of concern when deploying AI in the real world. The field of explainable AI was bolstered by the enforcement of the General Data Protection Regulation (GDPR) in the EU since 2018[5]. These regulations included the "right to explanation" from algorithm decisions, but also prohibited the processing of data that is unduly detrimental (i.e., unfair). Additionally, commercial providers of AI-

[*] SecDevOps: the process of integrating secure development best practices and methodologies into development and deployment processes which DevOps makes possible.

[†] MLOps: the process of integrating machine learning models into a continuous development production system.

based services have been struggling to gain acceptance after well publicized failures of AI-based services (e.g., AI services exhibiting gender and skin-type biases[6], AI recruiting tools that are biased against women[7], and IBM Watson providing "dangerous and useless" recommendations in healthcare settings[8]). These incidents have stoked general consumer anxieties over the widespread adoption of AI in many aspects of life and have driven organizations to seriously consider AI from the perspective of trust and ethics.

Despite this increased emphasis on trust, many topics remained to be explored. For example, the means to measure the AI uncertainty[9], the transferability of AI models to novel environments[10], data security, and realistic expectations about AI performance have only recently been emphasized as important considerations for deploying AI. On these fronts, limited public-private partnerships have begun to fill the gaps of research and development (e.g., the National Science Foundation's National Artificial Intelligence Research Institutes). Additionally, many academic groups (e.g., Stanford, Berkeley, MIT, and Carnegie Mellon) and non-profit organizations (e.g., The Future of Life Institute, and The Internet Society) have started new centers focusing on AI safety, explainable AI, and ethics.

Cohesive attempts at trusted AI, however, have largely been the domain of corporations. These include companies which sell turnkey AI solutions, such as IBM Watson and Microsoft AI Platform. Some corporate entities, such as Google, OpenAI, and DeepMind have put forth attempts at industry best practices.[11] While these efforts have grown in scale with commercial interest in safety critical applications (e.g., autonomous vehicles, medical AI, and cybersecurity) the impact outside of existing technology providers had been low.

Active AI Research Topics

Adversarial AI In 2016, researchers discovered that popular neural networks are susceptible to adversarial manipulation of inputs that cause them to provide erroneous predictions.

Explainable AI (XAI) With the popularity of highly complex, black-box AI algorithms, there has been growing concern over their lack of transparency. The field of XAI seeks to encourage the development of AI algorithms that can be understood and to create methods to illuminate the decisions of more complex AI systems.

Domain Adaptation With data and associated labels sometimes hard to come by, domain adaptation aims to leverage labeled data in one or more related source domains (e.g. synthetic data) to learn a ML model for unseen data in a target domain. This can be very challenging to accomplish in practice.

Uncertainty Quantification Any decisionmaking system requires both predictions and an associated uncertainty/confidence in that prediction. With ML models, especially deep learning (DL) models, predictions are sometimes both over-confident and wrong. If ML-based AI is ever to have a role in high consequence environments, this issue will have to be resolved.

Growing awareness of the benefits and risks of AI within all sectors of government has fueled a demand to not only deploy AI-based applications, but also verify that they can operate safely.[12,13,14,15] This has led to several investment research and development plans. As early as 2016, the Defense Advanced Research Projects Agency (DARPA) initiated the Explainable AI (XAI) effort to better

understand the predictions of AI algorithms. More recently, an update from the National Science and Technology Council on the national R&D AI strategy includes the goal of "creating robust and trustworthy AI systems."[16] In mid-2018, the DOD established the Joint Artificial Intelligence Center (JAIC) as a center of excellence with the mission to accelerate the adoption of AI for mission impact. In early 2020, the DOD adopted an official series of ethical principles for the use of AI.[17] As a result, the JAIC has paid increased attention to topics of trust within their Responsible AI Champions[18] program and the DOD Workforce AI Education strategy[19] that incorporates "responsible AI training" into multiple roles within the DOD.

The intelligence community (IC) has also released official strategy and guidance regarding the safety of AI systems. In the 2020 Artificial Intelligence Ethics Framework for the Intelligence Community and the 2020 Principles of AI Ethics for the Intelligence Community[20], the focus has shifted towards methods for designing in, testing, and measuring for trust. Most recently, the White House's National AI Initiative Office released a list of the Characteristics of Trust in January 2021.

Based on the above government-led initiatives to increase understanding and trust in AI, many federally funded research and development centers (FFRDCs) have developed doctrine for considering trust (e.g., MITRE's Trust in Autonomous Systems and the Institute for Defense Analysis's Roadmap to Assurance) and commercial entities have also responded with customer-focused strategies (e.g., Deloitte's Trusted AI Framework). Going forward, the need to provide tailored guidance for the deployment of AI in the space domain will increase. Many opportunities are present to shape AI policy in this challenging environment. Collaboration between research groups (such as universities and FFRDCs) could bring the insights gained in the academic world directly to government customers. An objective collaborative approach free of vested commercial equities could mitigate future potential for vendor lock and commercial hegemony over cutting-edge AI knowledge.

Existing Policy

In recent years, policies related to the ethical considerations of AI have emerged. Of note, the European Commission appointed the High-level Expert Group of AI (AI HLEG) to develop a long-term strategy on AI development and establish ethical priorities. In April 2019, the AI HLEG produced the "Ethics Guidelines for Trustworthy AI," which recommended a set of requirements that AI systems should meet to be deemed trustworthy.[21] While providing a strong set of standards to frame AI development, responses to the AI HLEG have not all been positive, with some policy papers suggesting that AI can never be fully trusted.[22,23,24]

Within the United States there has been considerable attention at various levels of government to encourage the research and development of trustworthy AI and AI more generally. In early 2019, the White House released executive orders that outlined U.S. policy to sustain and enhance the leadership position of the U.S. in AI research and development.[25] In late 2020, this executive order was expanded to include promotion of the use of trustworthy AI in the federal government. This emphasis on AI strategy has also created international partnerships, such as a recent cooperative effort between the U.S. and UK to advance the development of trustworthy AI.[26]

As a result of recent executive orders within the U.S., subsequent policies have been established in other areas of government. In particular, the Directory of the Office of Management and Budget (OMB) published the *Guidance for Regulation of Artificial Intelligence Applications*, which included sections emphasizing the need to build public trust in AI.[27]

A Framework for Trusted AI

Within high consequence environments, the criteria for trust are stringent and multidimensional. Stakeholders at multiple levels of program management and at various points within project development will want to minimize the risk of failing to meet project goals and objectives. When AI algorithms are proposed as new capabilities or as enhancements to existing ones, the framework for trusted AI will help to build confidence that risks can be identified, quantified, and mitigated as best as possible.

Policymakers can encourage the incorporation of trust into the development of AI-based capabilities. This would shape project development from the earliest stages by pairing performance expectations with the need to verify that performance expectations have been met. As a result, project requirements would include additional or modified objectives that accommodated verification of trust in AI-based algorithms. Ideally, these requirements would include reference to attributes of trust, which are responsible for measuring and tracking properties of AI behavior. Finally, trust will need to be verified not only in the component AI algorithms, but also in monitoring and safeguarding tools. These concepts are formalized into a framework by defining three threads, each of which offers a different perspective on how to measure trust. These threads also map to a project development lifecycle to facilitate its incorporation at key points.

Thread 1: Objective and Data Specification

The first thread is focused on answering the question *What is the task and how will data be acquired?* and maps to the first two steps in the AI lifecycle above. It assumes that there is a need and justification for employing AI, based either on an identified capability gap or the need for capability enhancement. After the need is established, the objectives, constraints, and any limitations of the AI should be identified. A plan for how to collect, prepare, and characterize data is also essential since data is instrumental in the implementation and evaluation of AI algorithms. This thread sets the

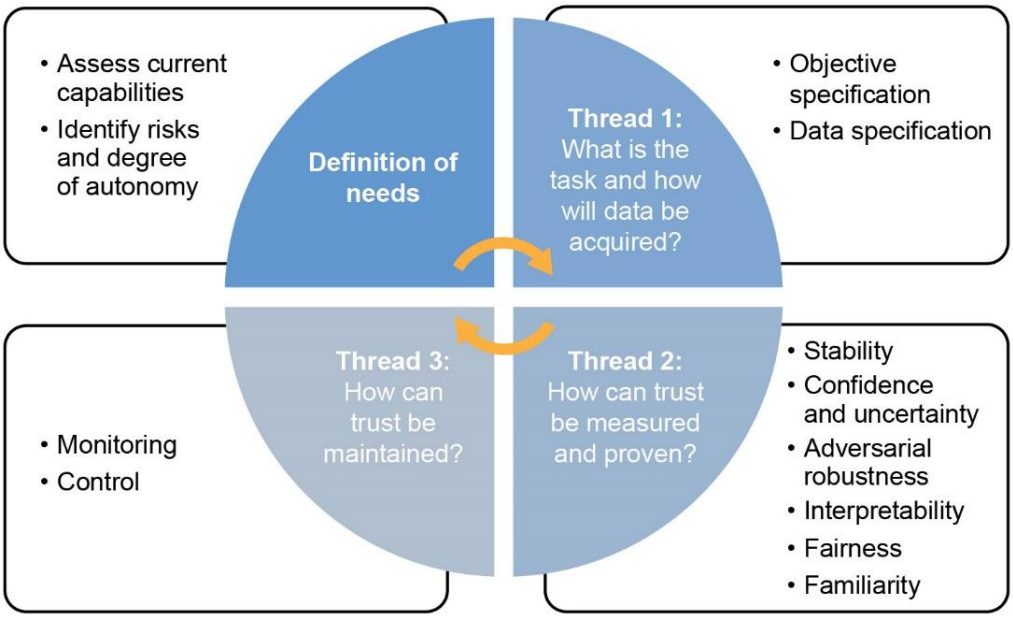


Figure 1: The threads of Trusted AI as they are applied within the development lifecycle. As AI is deployed in real-world environments, this cycle can repeat as new needs or AI performance limitations emerge.

stage for subsequent steps as it defines what is expected from the AI and how it will be deployed within a larger system. It is broken into two stages: objective specification and data specification.

Objective Specification

An objective specification is a clear description of what task the AI will perform in its deployed state along with a plan for how any AI-based algorithms will be assessed. It should be produced in conjunction with subject matter experts (SMEs) who understand mission needs and AI developers who can translate those needs into an objective that can be accomplished by an AI. Both are needed as SMEs may not know what is feasible with current AI algorithms or the terminology to describe their objective within the domain of AI. Therefore, it is up to both SMEs and AI developers to understand mission objectives and articulate how AI can be leveraged to meet them.

As motivation for this, there are many examples in AI literature of algorithms that were trained to accomplish an objective but ended up doing so in unexpected ways. This is often due to the difficulty in translating a user-specified task into an objective that an AI can learn to accomplish.[28] Some notable issues include the following:

- Real world use cases can involve a complex environment with many possible states and actions that an AI must contend with. It will be impossible to train against all possible events and outcomes that could lead to issues during deployment. To mitigate this, an objective specification should include not just the desired behaviors, but also a set of known failure modes and unallowable state conditions.

- Objectives are stated in subjective language or represent highly complex behavior. This creates issues when trying to develop an AI that accurately represents the desired need. Therefore, the objective should be broken down into manageable and measurable sub-tasks that collectively represent the desired behavior.

- There is significant risk of harm to the external environment or users. In those settings, there should be plans to implement safeguards or severe penalties against actions that would impact the operational environment of an AI. These should also serve as components within a monitoring tool, which is further discussed in Thread 3: Monitoring and Control.

Data Specification

The data specification is comprised of a plan for creating a dataset, along with the means to sample and prepare data for both training and evaluation. This will help to determine whether adequate data are available for training an AI, whether domain adaptation will be a concern (e.g., when synthetic data will be used to train a ML model), and anticipate how well the trained AI will succeed when deployed and exposed to its operational environment.

A known issue with AI algorithms, and especially ML models, is that the deployed performance is significantly worse than what was observed during training. Understanding the cause of this can be challenging, but is often due to poor assumptions made during the selection of training data. Even worse, the deployed performance may be unknown since ground truth information may be unavailable for AI assessment. The following steps can help to mitigate this:

- Meta-information related to any data collection can help to identify issues before they occur in deployment. A data specification should include any seasonal variations (i.e., time-dependent effects that will have to be captured), the source of collected data (e.g., the type of sensor and its characteristics), any data cleaning steps, and pre-processing done prior to AI training.

- Knowledge of what are routine and exceptional data properties. The data specification must include statistical characterization of data collected from sources (even if that source is a synthetic data generator) and when used in model training. This will help inform attributes of trust discussed in Thread 2: Trusted AI Attributes.

Thread 2: Trusted AI Attributes

After the AI objectives and data properties have been described, there is still a need to assess how well an AI-based application can be trusted during its development and deployment. Trust can and should be evaluated from multiple viewpoints. For example, some metrics should emphasize model performance on challenging data, while others will need to capture how comfortable human operators are using AI in their jobs. The attributes collected below comprise a set of broadly applicable approaches to evaluating the level of trust:

- **Stability**: Establishing trust in an AI-based application starts with meeting basic assumptions and expectations about how that application will perform and how stable that performance is on routine data inputs. *Stability is the consistency of model performance when provided inputs that fall within a routine range of data parameters*. Two evaluation methods are recommended: third party verification and pre-deployment testing. Third party verification involves providing an original data source, the specification for how the data were collected and prepared, expected performance metrics, and any additional details necessary for an independent group to replicate the expected performance.[29] A pre-deployment verification effort places the AI within an environment that replicates, as accurately as possible, the deployment environment and the data inputs that the AI will encounter.

- **Confidence**: As discussed above, the predictions made from AI algorithms can be both incorrect and overconfident.[30] Confidence is quantification of the sureness of the model output across the entire input space that is consistent with the observed error rate. This confidence should be high for inputs that closely match routine inputs (and few observed errors) and low for exceptional inputs (where more errors can occur). Whenever possible, prediction intervals should be provided to bound confidences.

- **Uncertainty**: Many AI algorithms are not able to provide sensible outputs when presented with novel inputs but should at least be able to notify users that their predictions should not be trusted. *Uncertainty in AI is the ability to discern when inputs fall within exceptional ranges of the data distribution and provide bounds for when AI predictions should have low confidence*. Typically, an auxiliary technique is required to detect when an input falls within an exceptional range, with some successful approaches demonstrated on high dimensional data.[31]

- **Adversarial Robustness**: It is well established in both academic and mainstream press that any AI algorithm trained from data is likely susceptible to adversarial attacks.[32] These are easily accomplished by modifying input data in such a way as to confuse the AI algorithm. *Robustness in the context of adversarial attacks is defined as the AI's ability to provide outputs consistent with inputs when no attacks are present and to detect when an attack has occurred*. Consistency can be assessed in two ways: 1) Comparing the effects of perturbed inputs and unperturbed inputs on AI algorithm predictions and their associated attributions

and/or 2) Assessing how much AI algorithm predictions and attributions have changed on unperturbed inputs after a poisoning attack has occurred.

- **Interpretability**: As the complexity of recent AI algorithms has increased, they have also become notorious for being opaque black boxes. This has raised concerns in fields where AI is meant to interface closely with users (e.g., medical AI) and some degree of interpretability is critical for building trust. *Interpretability is defined as the degree to which a user can understand the cause of an AI algorithm prediction.* This goes beyond explainabilty, which simply requires the availability of attributions, and demands that those explanations must also reduce the burden of user comprehension. Developers should incorporate two mechanisms: attributions that indicate how data influenced model predictions, and the means to predict the utility of those attributions. The utility of an attribution, or explanation, is what determines interpretability and should be developed with user input.

- **Fairness**: When making predictions or decisions which impact users or the external environment, trust is quickly lost when AI predictions are inconsistent between users and different contexts. *Fairness is defined as providing equitable outcomes to all subsets of the population or environment*. Whenever data is used to train an AI, any biases present in the data will be relearned and reinforced unless efforts are taken to take those biases into account. To mitigate this concern, developers should analyze data for any unbalanced representations of data subgroups, look for inconsistent performance between subgroups, and incorporate any relevant data imbalance mitigation strategies.[33]

- **Familiarity**: Users will be more likely to trust an AI if they are familiar with under what conditions it performs well. *Familiarity is defined as users being able to anticipate the predictions of an AI algorithm*. However, in many real-world scenarios, this is difficult to achieve. This can be addressed by developing AI-based applications from well-understood algorithms and/or datasets. When feasible, familiarity can also be garnered by operating the AI in "shadow mode" (where it generates predictions without directly impacting decisions) or by gradually increasing the degree of risk of its deployment (e.g., by first deploying it to perform an auxiliary task and then using it in more critical settings).

Thread 3: Assuring Deployed Model Maintains Attributes of Trust

The final thread maps directly to the final phase of the AI lifecycle, which is to consistently evaluate that an AI-enabled system maintains the attributes of trust while deployed to its operational environment. If the model does not maintain trust during its time in operations, then the lifecycle—and the defined threads—cycles back to the start of the process. This point in the cycle indicates that the AI should be updated (or a new one created). To facilitate this process, two mechanisms are recommended: monitoring of AI to support assessment of the attributes of trust and some degree of control to interrupt AI operation if something goes wrong.[34]

- **Monitoring**: *Monitoring is the automated and continuously available assessment of AI-based applications to verify the attributes of trust are maintained after deployment*. This is facilitated through several development steps of trusted AI development: 1) The implementation of sub-processes to measure physical constraints and avoid failure modes, as defined in the objective specification, 2) Definition of routine and exceptional inputs, along with expected data properties, from the data specification, 3) Expected performance metrics, 4) Confidence and uncertainty predictions, and 5) Detection of

adversarial attacks. These metrics can be collected continuously and inform a high-level assessment of application health.

- **Control**: *Control is the ability to interrupt or terminate AI execution when undesirable behavior occurs and to do so with minimal impact on other systems*. Multiple levels of control could be included in the application, which would be used under different scenarios. The specific controls for a system should depend on the operational environment of the application, degree of risk, and access to users for possible intervention. These control methods should be tested as part of system or architectural-level testing to ensure unexpected effects are not propagated beyond the AI-based application.

The Road Ahead

The framework described above highlights the essential components to establishing trust by providing clearly defined metrics that measure how well trust has been satisfied. However, research into many of these concepts is ongoing and some are not at a level of maturity appropriate for AI deployment. Therefore, roadmaps for future research and significant investments will be needed to further enhance and understand trusted AI. This will be particularly true as AI is applied to more diverse settings and environments and the operational needs continue to evolve. Regardless, the threads of the Trusted AI Framework provide the starting point for policymakers to appreciate the current limitations of AI and where additional attention and resources are needed.

These threads represent aspects of AI development that must be considered across a range of applications. However, many of these will require a different emphasis on concepts within the threads or altogether new ones. Some applications will have a strong emphasis on security and privacy. In those cases, trust would likely include concepts that strive to guarantee data privacy and protection. Other applications will require frequent cooperation with users, focusing on human-machine teaming, while others will be completely autonomous, such as those operating in remote environments. Further applications will require careful considerations of how algorithms trained in a lab can be translated to deployable software and hardware environments. In the pLEO example discussed earlier, trust will depend to a large degree on verifying that complex autonomous behavior is still reliable and safe with minimal manual intervention. These examples demonstrate that attempts to collect all aspects of trust into a single framework will never capture all relevant concepts for all applications. Future research in Trusted AI should focus on extending the threads of trusted AI to specific application areas and highlighting the challenges in each.

Impact on Future Policy

As described earlier, determining whether AI is appropriate is based on three distinct cases: 1) There is a need that is not met by current capabilities and that can only be enabled by AI, 2) AI offers a potential enhancement to existing capabilities, or 3) AI has already been deployed within an operational system and trust must be established post-deployment. In situations where there is a need for new capabilities enabled by AI, a framework can help provide proof that an AI-based algorithm can be trusted and deployed. For example, in fully autonomous space missions there may be reluctance to employ a new AI capability when simpler preexisting methods may suffice. By utilizing the attributes of trust, policymakers can encourage people to gather the right evidence to not only assess the performance of an AI-based algorithm, but also faithfully compare the risks and benefits of deploying AI-based algorithms instead of relying on existing less capable methods. Second, in cases where AI could offer enhancements to existing

capabilities, the framework provides policymakers the motivation to urge careful consideration within their domain. As discussed throughout this paper, there are many known pitfalls of AI-based algorithms and significant effort is required to mitigate their impact on real-world systems. As a result, policymakers could use the framework to gauge the additional level of investment required for enhancing systems with AI in high consequence environments. Finally, in cases where AI has already been deployed, the framework provides the groundwork for developing metrics and tools to prove to stakeholders that the level of trust in AI-based systems can be systematically understood.

Conclusion

The need for trust in AI-based applications is paramount in high consequence environments. Whenever applications operate completely autonomously, in conjunction with humans or in situations that potentially have significant impact on an external environment, a set of best practices must be in place to minimize the chance of adverse consequences. This document provides one set of best practices and collects them into a framework that covers the lifecycle of AI development. This framework will not only help evaluate trust in deployed AI, but also help policymakers refine their vision for when and how to deploy AI. As it becomes increasingly apparent that AI will touch every aspect of our daily lives, the Threads of Trust of the Trusted AI Framework will help policymakers have the confidence to pursue the benefits of AI while also ensuring the AI-enabled future.

Acknowledgments

The authors would like to express their gratitude to Mike Tanzillo, Brian Hardt, Dorothy Arbiter, Susan Herbulock, Marcus Stefanou, Mike Nemerouf, Josef Koller, Jamie Morin, Karen Jones, Russell Rumbaugh, Robin Dickey, Amy O'Brien, Ron Birk, and Zigmond Leszczynski for their helpful reviews and comments.

References

[1] Marcus, Gary and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon, 2019.

[2] SecDevOps: the process of integrating secure development best practices and methodologies into development and deployment processes which DevOps makes possible.

[3] Defense Advanced Research Projects Agency. (August 2016). Explainable Artificial Intelligence (XAI) (https://www.darpa.mil/program/explainable-artificial-intelligence).

[4] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." arXiv preprint arXiv:1611.01236 (2016).

[5] Goddard, Michelle, "The EU General Data Protection Regulation (GDPR): European regulation that has a global impact," International Journal of Market Research 59.6 (2017): pp. 703-705.

[6] Hardesty, Larry. "Study finds gender and skin-type bias in commercial artificial-intelligence systems." Retrieved April 3 (2018): 2019.

[7] Dastin, Jeffrey (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G).

[8] Strickland, Eliza. "IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care." IEEE Spectrum 56.4 (2019): pp. 24-31.

[9] Begoli, Edmon, Tanmoy Bhattacharya, and Dimitri Kusnezov. "The need for uncertainty quantification in machine-assisted medical decision making." Nature Machine Intelligence 1.1 (2019): pp. 20-23.

[10] Csurka, Gabriela. "Domain adaptation for visual applications: A comprehensive survey." arXiv preprint arXiv:1702.05374 (2017).

[11] Perspectives on Issues in AI Governance (https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf).

[12] Porter, Daniel, McAnally, Michael, Beiber, Chad, Wojton, Heather, and Medlin, Rebecca (2020, May). Trustworthy Autonomy: A Roadmap to Assurance, Part 1: System Effectiveness (IDA document: P-10768). Institute for Defense Analyses (https://www.ida.org/research-and-publications/publications/all/t/tr/trustworthy-autonomy-a-roadmap-to-assurance-part-1-system-effectiveness).

[13] Trustworthy AI. (2020, August 26). Deloitte United States (https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html).

[14] Trusting AI: IBM Research. (2018) (https://www.Research.Ibm.Com/Artificial-Intelligence/Trusted-Ai/. https://www.research.ibm.com/artificial-intelligence/trusted-ai/).

[15] Artificial Intelligence (https://www.nist.gov/artificial-intelligence).

[16] Select Committee on Artificial Intelligence of the National Science and Technology Council. (June 2019). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update.

[17] United States Department of Defense. (February 24, 2020). DOD Adopts Ethical Principles for Artificial Intelligence (https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/).

[18] DOD Joint AI Center. (August 2020). Department of Defense Joint Artificial Intelligence Center Responsible AI Champions Pilot.

[19] DOD Joint AI Center. (September 2020). 2020 Department of Defense Artificial Intelligence Education Strategy.

[20] ODNI. (2020, June). Artificial Intelligence Ethics Framework for the Intelligence Community. Office of the Director of National Intelligence (https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community).

[21] Floridi, Luciano. "Establishing the rules for building trustworthy AI." Nature Machine Intelligence 1.6 (2019): pp. 261-262.

[22] Bryson, Joanna. “AI & Global Governance: No One Should Trust AI.” United Nations University: Center for Policy Research. November 13, 2018.

[23] Fjelland, Ragnar. "Why general artificial intelligence will not be realized." Humanities and Social Sciences Communications 7.1 (2020): pp. 1-9.

[24] Ryan, Mark. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." Science and Engineering Ethics 26.5 (2020): pp. 2749-2767.

[25] Executive Order 13859. (February 11, 2019). Maintaining American Leadership in Artificial Intelligence.

[26] Bureau of Oceans and International Environment and Scientific Affairs. (2020, September). Declaration of the United States of America and the United Kingdom of Great Britain and Northern Ireland on Cooperation in Artificial Intelligence Research and Development: A Shared Vision for Driving

Technological Breakthroughs in Artificial Intelligence.

[27] Director of the Office of Management and Budget. (January 2020). *Guidance for Regulation of Artificial Intelligence Applications*.

[28] Leike, Jan, et al. "Scalable agent alignment via reward modeling: a research direction." arXiv preprint arXiv:1811.07871 (2018).

[29] Pineau, Joelle (2020, April 7). The Machine Learning Reproducibility Checklist. McGill (https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf).

[30] Guo, Chuan, Pleiss, Geoff, Sun, Yu, and Weinberger, Kilian (2017). On Calibration of Modern Neural Networks. 34th International Conference on Machine Learning, Sydney, Australia.

[31] Papernot, Nicolas and McDaniel, Patrick (2018, March 13). Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. ArXiv.Org (https://arxiv.org/abs/1803.04765v1).

[32] Danks, David (2020, February 26). How Adversarial Attacks Could Destabilize Military AI Systems. IEEE Spectrum: Technology, Engineering, and Science News (https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/adversarial-attacks-and-ai-systems).

[33] Johnson, Justin M., and Khoshgoftaar, Taghi M. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): pp. 1-54.

[34] Ortega, Pedro, et al. Building safe artificial intelligence: specification, robustness, and assurance. 2018.

AEROSPACE REPORT NO.
ATR-2021-01456

A Framework for Trusted Artificial Intelligence in High-Consequence Environments

17 May 2021

Philip C. Slingerland[1], Lauren H. Perry[2]
[1]Machine Intelligence and Exploration Department, RS Signals and Analytics Division
[2]Space Application Group, Survivability and Resilience Department

Prepared for: Senior Vice President, Engineering and Technology Group

Authorized by: Engineering and Technology Group

Revision History

Date	Rev	
April 2020	TOR Rev -	Initial Release
September 2020	TOR Rev A	Incorporated new relevant publications, Updated attribute opportunities, Added SceptreML illustration
May 2021	ATR Rev -	Incorporated new relevant publications, Added Risk to Mission Integrity concept Changed Reproducibility to Stability Modified attribute opportunities to implementation alternatives

Motivation

- Trust and safety of AI is a nascent, but rapidly growing field
  - *Meaning and importance of "trust" differs depending on the application*
  - *Definitions of trust are slowly converging within the context of AI/ML-enabled systems*
- Trust is a suitcase word. Suitcase words, as described by M. Minsky (cognitive scientist with focus on AI and co-founder of MIT Computer Science and AI Labaratory), are
  - *"Words that all of us use to encapsulate our jumbled ideas about our minds. We use those words as suitcases in which to contain all sorts of mysteries that we can't yet explain…Inside that suitcase are assortments of things whose distinctions and differences are confused by our giving them all the same name." [1]*
  - *"Words we all recognize and understand but have a hard time explaining, such as emotions, consciousness, and thinking…they contain many smaller concepts that can be unpacked and analyzed" [1]*
- The goal is to break down aspects of "trust" of an AI/ML-enabled system into a set of meaningful, generalizable, measurable and testable attributes
  - *The need for trust and how trust should be assessed can vary widely among different domains, applications and level of autonomy of the system*

How can trust be defined and quantified within an AI/ML-enabled system?

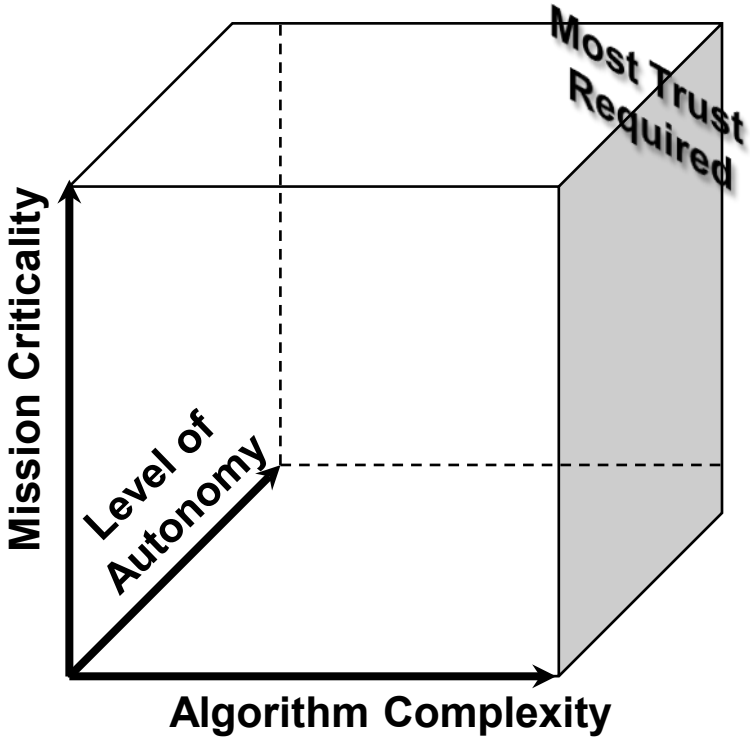
Why Trust Matters in Safety Critical Systems

- The growth of ubiquitous AI, of which many applications are mission critical, is driving the need for AI systems to be trusted to an increasingly higher degree.
- While there are known limitations to current AI algorithms, their ability to enhance current systems and enable new capabilities is well-established
- The barrier to adoption of these systems is not always algorithm behavior, but rather the requirement to quantify and bound risk from program managers:
  - *AI-enabled systems may be held to a higher standard than human decision-makers or action officers. Engineers are required to demonstrate that their AI algorithm will accomplish the task, as well as how. [2]*
  - *Hesitancy to utilize algorithms that have minimal heritage or no proven performance in a mission critical context*
  - *Additional complexity of AI algorithms means increased cost to develop, test, validate, and monitor the safe operation of those algorithms*
  - *With little insight into how an AI algorithm functions, establishing trust of an AI algorithm for use in an autonomous system is difficult to achieve*
- Aerospace's Trusted AI framework can help with understanding and mitigatation of the risks of incorporating AI
  - *It requires engineers to clearly demonstrate how their AI algorithm will accomplish a task. This will help program managers understand how the algorithm will operate and what new capabilities it will bring.*
  - *By providing best practices for how to measure trust, program managers can plan and budget for sufficient development of AI algorithms.*

The framework can help programs leverage AI by better understanding risk

Understanding the Amount of Trust Required

AI Risk Cube



- The amount of trust required is directly related to "Risk to Mission Integrity" — a metric which can be defined by three dimensions:
- **Mission Criticality**
  - Trusted AI initiatives will assist with properly assessing the risk to mission success when AI is applied, based on the specific function for which the algorithm is meant to accomplish
- **Algorithm Complexity**
  - Disciplined approach to AI development will assist with managing algorithmic complexity
- **Level of Autonomy**
  - Autonomy should be applied purposefully to serve as a force multiplier to maximize user/opportunity, efficiency, and capability

The level of trust required will vary depending on the combination of Algorithm Complexity, Mission Criticality, and Level of Autonomy

Approach To Defining Trust

- Investigated efforts in Trusted AI/ML across commercial, government, and academic organizations (October – November 2019)
- Met with Aerospace AI/ML SMEs to discuss perspectives on what is needed to trust AI/ML-enabled systems in customer applications (November 2019)
- Performed literature review to understand state-of-the-field (November 2019 – December 2019)
- Generalized external scan terminology and approaches to increasing trust in AI applications to develop a set of Trusted AI threads (January 2020)
  - *Threads are set of themes of how to better understand, test, and monitor the AI/ML algorithms being developed so users can gain and maintain trust of the system*
  - *Trusted AI threads are applicable to both data-driven AI and model-driven AI, however examples are focused on customer-related, data-driven AI concerns*
- Trusted AI threads are continually updated as The Aerospace Corporation funds internal studies that develop and implement thread attributes while also following ongoing external research efforts (January 2020+).

Insights From 2019 Trusted AI Literature Review

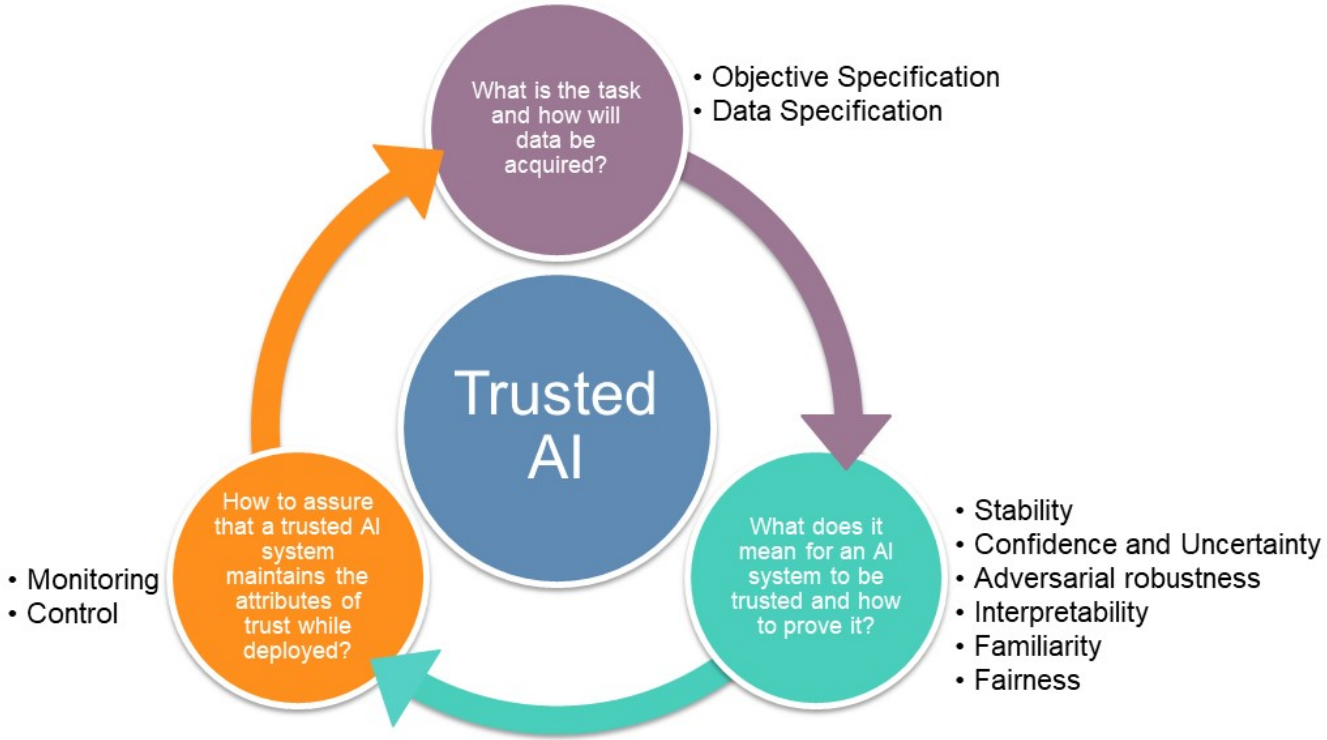
- Until recently, most research has focused almost exclusively on Adversarial and Explainable AI. What changed?
  - *Enforcement of General Data Protection Regulation (GDPR) in the EU since 2018 requires not only "right to explanation" from algorithmic decisions, but also prohibits processing data that is unduly detrimental (i.e., unfair)*
  - *Highly publicized examples of AI bias and failures have stoked anxieties over the widespread adoption of AI in all aspects of life. This has forced organizations to seriously consider AI from perspectives of trust and ethics.*
- Most organizations focus on an individual problem
  - *Multiple public-private partnerships concentrate on specific aspects of trusted AI or assured intelligent autonomy*
  - *Some larger organizations (such as Microsoft and IBM) and government organizations (such as NIST) are researching generalizations*
- University research is rapidly expanding in this area but work often has minimal overlap with safety-critical applications in defense and intelligence.
  - *Since 2017, Stanford, Berkeley, Carnegie Mellon, and other institutions have started new centers focusing on AI safety, explainability, and ethics*

Definitions of Trust Across Industry and Government

- Around the same time as the initial publication of Aerospace's Trusted AI Framework (Apr 2020), similar frameworks were also published:
  - *IBM's Trusting AI Focus Areas (2019)*
  - *Department of Defense's Ethical Development of AI Capabilities (Feb 2020)*
  - *Deloitte's Trusted AI Framework (Mar 2020)*
  - *IDA Roadmap to Assurance (May 2020)*
  - *Artificial Intelligence Ethical Framework for the Intelligence Community (July 2020)*
  - *NIST Workshop on AI Trustworthiness (Aug 2020)*
  - *Microsoft Principles of Responsible AI (Jan 2021)*
  - *National AI Initiative Office's Characteristics of Trust (Feb 2021)*

Aerospace Trusted AI Framework

- We define **Trusted AI** as having *actionable confidence that the AI algorithm and its characteristics meet user defined objectives in a proper and understandable way over the lifetime of the system*
- The three threads of trusted AI are a set of recommended best practices to demonstrate trust



- With these questions in mind from the start, the trust of an AI-enabled system can be achieved
  - *Requires investments in time and attention*
  - *Acceptance and buy-in from AI practitioners is critical*
- If the model does not maintain trust during its time in operations, then the lifecycle — and thus the defined threads — cycle back to the start, as the model should be updated (or a new model created)

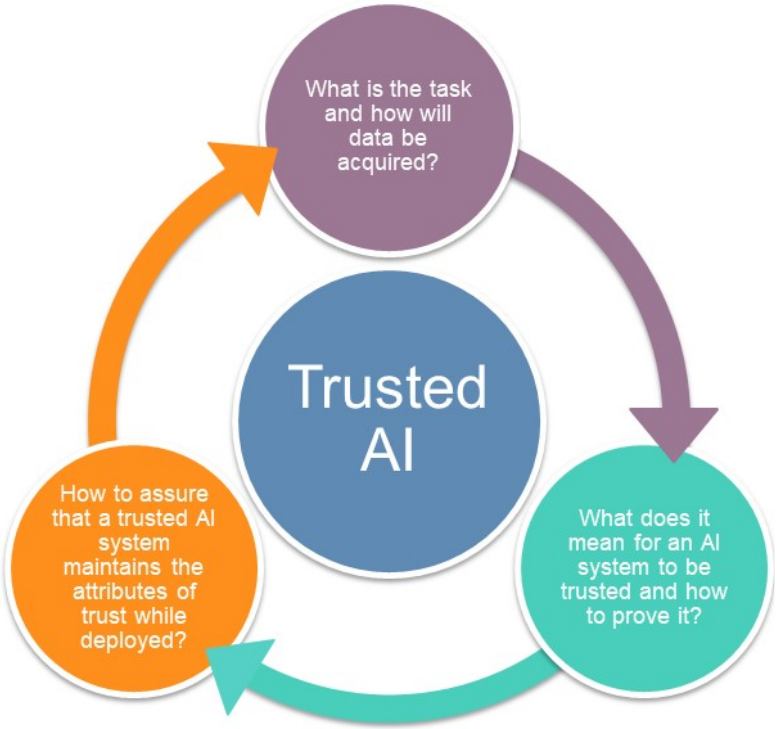
Trusted AI is as much a philosophy and engineering process as it is a system feature

Comparing Different Definitions of Trust

Aerospace's Trusted AI Framework		National AI Initiative Office Characteristics of Trust	DoD's Principles of AI Ethics		AI Ethics Framework for the Intelligence Community		Deloitte's Trustworthy AI framework	IBM - Trusting AI		Microsoft Responsible and Trusted AI
Thread 1	Objective Specification	Ethical use of AI	Responsible, Traceable, Reliable		Testing, Version Control (builds, models, data), Stewardship	Governing the AI and the data; Documentation of Purpose, Parameters, Limitations, and Design Outcomes		Transparency and Accountability	Value Alignment	
	Data Specification	Privacy					Privacy			Privacy and Security
Thread 2	Stability	Accuracy, Reliability					Robust/Reliable			
	Confidence and Uncertainty									
	Adversarial Robustness	Robustness, Resilient					Robust/Reliable		Robustness	
	Interpretability	Explainability and Interpretability, Transparency				Transparency: Explainability and Interpretability	Transparent/ Explainable		Explainability	Transparency
	Familiarity									
	Fairness	Fairness, Bias Mitigation		Equitable		Mitigating Undesired Bias and Ensuring Objectivity	Fair/Impartial		Fairness	Fairness, Inclusiveness
Thread 3	Monitoring	Security		Governable		Periodic Review	Safe/Secure	Transparency and Accountability		Reliability and Safety
	Control	Safety				Human Judgement and Accountability	Responsible/ Accountable			Accountability

Aerospace's Trusted AI Framework encompasses the focus areas of several trust frameworks, while providing explicit guidance on how to accomplish trust in relevant applications

The Threads of Trusted AI



Thread 1: What is the task?

Objective Specification

- **Problem Statement:** An AI algorithm can learn to exploit a poorly specified objective or a flaw in the training environment to give the false impression that it has "learned" to accomplish a task. To minimize the risk of deploying an improperly trained AI, users must ensure the objective was accomplished in a manner consistent with user need and expectations.

- **Example:** Satellite agent is given the objective to maximize a number of collected images. The agent de-emphasizes collects far from its current pointing vector, as collect priority was not added as part of the objective function.

- **Description:** Challenges arise not only in defining an objective, but in translating it into a set of functions that an AI can optimize against
  - *A trusted AI system must have precise definitions for both the user-specified objective and the objective accomplished by the AI, to enable quantification of their agreement.* [3,4]
  - *Objectives should include the expected AI performance metrics. This will guide bias/variance, interpretable/black box tradeoffs that will occur during training and deployment.* [5]
  - *AI training requires detailed knowledge of both the task and how well the AI is adhering to the original intent of the specified objective* [6]
  - *A clearly defined objective supports reproducibility of results and independent algorithm validation*
  - *Identifying what data is required to accomplish the objective, or if data of a suitable quality can be obtained. (Garbage-In, Garbage Out is still applicable to AI algorithms).*
- **Implementation Alternatives:**
  - *Adopting formal methods for defining the objectives of an AI algorithm in a way that can be engineered against and compared (Aerospace, 2021)*

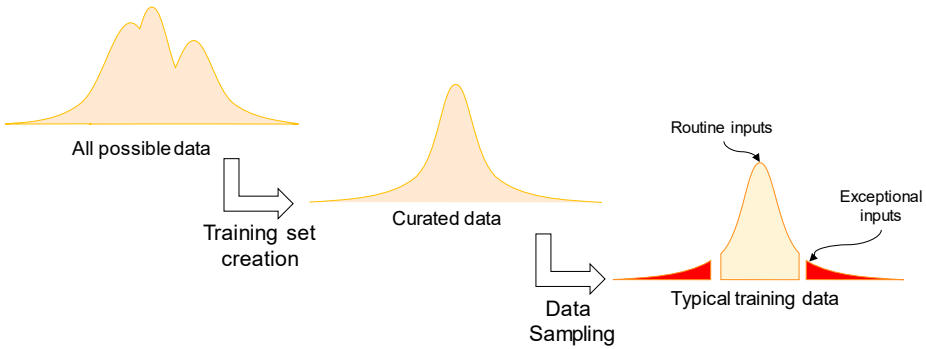
Objective Specification will provide the groundwork for defining the standard by which the AI will be assessed

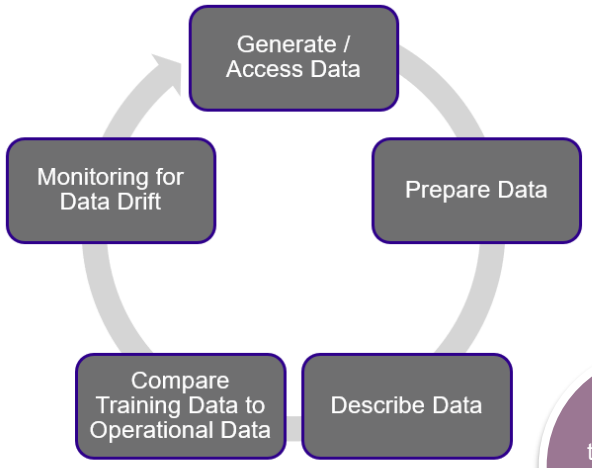
What is the task and how will data be acquired?

Thread 1: How will data be acquired?

Data Specification

- **Problem Statement:** Performance of deployed AI can be significantly worse than expected once encountering real data in an operational environment due to noise or other factors
- **Example:** Data labeled for training a machine learning algorithm on remote sensing task only contained images with no clouds, thus the deployed system is biased to only perform well on cloud-free images
- **Description:** Assumptions made during selection of training data must be understood to ensure accurate representation of deployed environment data (selection bias, population shifts, sensor characteristics, etc.) [7]
  - *Specify and articulate data collection process to prevent biases which may affect deployed AI performance* [8]
  - *Data specifications can help define boundary between algorithmic routine and exceptional inputs*
  - *Data specifications support monitoring for data drift to alert when an algorithm needs to be retrained*
- **Implementation Alternatives:**
  - *Quality of AI Data Checklist* [9]
  - *Training data configuration management using MLOps*
  - *Quantify domain transfer effects from simulated to real data*





Data Lifecycle

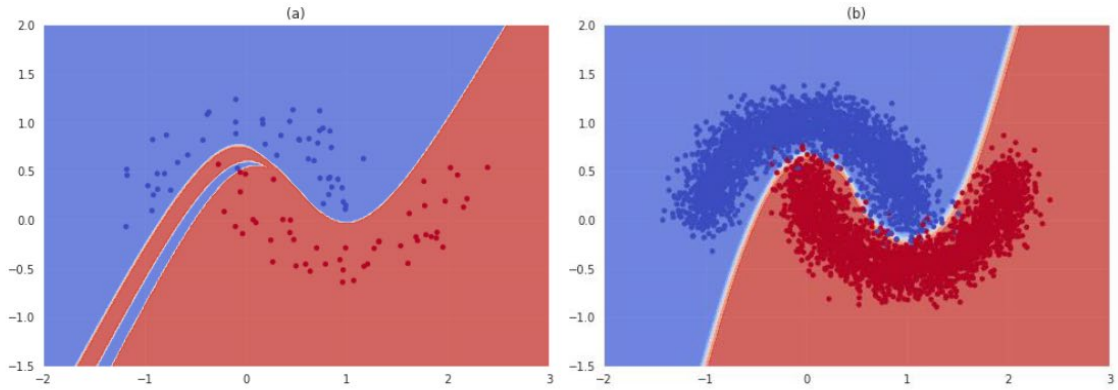
What is the task and how will data be acquired?

Trusted AI requires specifying processes of training set creation and sampling

Thread 2: Trusted AI Attributes and Metrics

Stability

- **Problem Statement:** Deployed AI may not always provide consistent or similar responses to similar inputs or even inputs that appear identical to the human eye
- **Example:** Due to sample biases and/or inadequate data variation present during model training, a deployed model may be improperly sensitive to input parameters and performs inconsistently and/or unpredictably when encountering operational data
- **Description:** Stability is the consistency of model predictions when provided inputs that fall within a routine range of data parameters
- **Implementation Alternatives:**
- Google's Robustness Metrics (https://github.com/google-research/robustness_metrics)
    - Out-of-distribution generalization
    - Stability under natural input perturbations



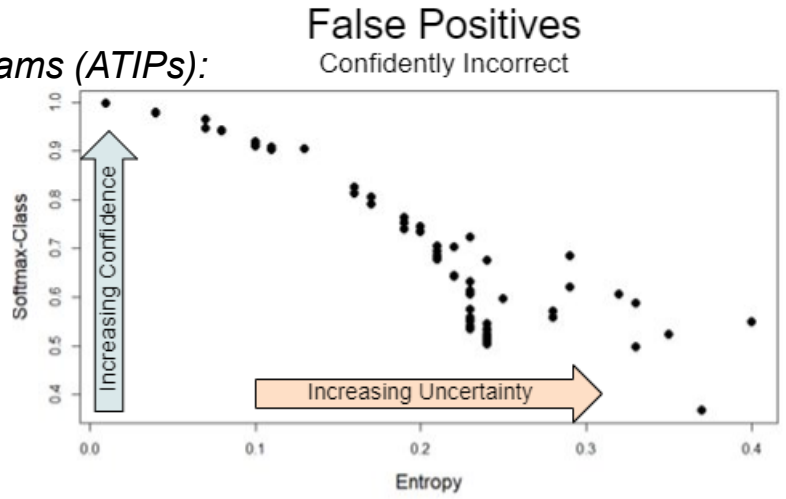
Training data and decision boundaries from two training runs using different sample sizes. (a) Sample size of 100 resulted in overly complex decision boundary. (b) Sample size of 5000 resulted in simpler, but more accurate decision boundary.

What does it mean for an AI system to be trusted and how to prove it?

Thread 2: Trusted AI Attributes and Metrics

Confidence and Uncertainty

- **Problem Statement:** Many AI algorithms provide highly confident but incorrect predictions, especially on data that occur in rare, unexpected, or novel environments. However, in our domain it is the rare and unexpected events that are often of most significance to us.
- **Example:** Automatic target recognition (ATR) algorithm that detects and classifies aircraft by manufacturer was originally trained using satellite imagery of North American airports. When deployed globally, the algorithm should demonstrate reduced confidence of a prediction when observing aircraft from rare or never-before-seen manufacturers.
- **Description:** Confidence is the quantification of the sureness of the model output across entire the input space and should be calibrated to match the model performance. Uncertainty is the ability to discern when inputs fall within unexpected or exceptional ranges of the input space to provide bounds for when model outputs will be unreliable.
- **Implementation Alternatives:**
  - *Monte-Carlo Dropout for Quantifying and Leveraging Prediction Uncertainty (Aero AI CSI funding 2020 and 2021*,[10]*)*
  - *Aerospace Technical Improvement Programs (ATIPs):*
    - Prediction intervals for Neural Networks (2020)
    - Deep Ensembles for Uncertainty Quantification (2021)
    - Auto-Encoder Out-of-Distribution Testing (2021)
    - Expected Calibration Error (2021)
    - Reliability Diagrams (2021)

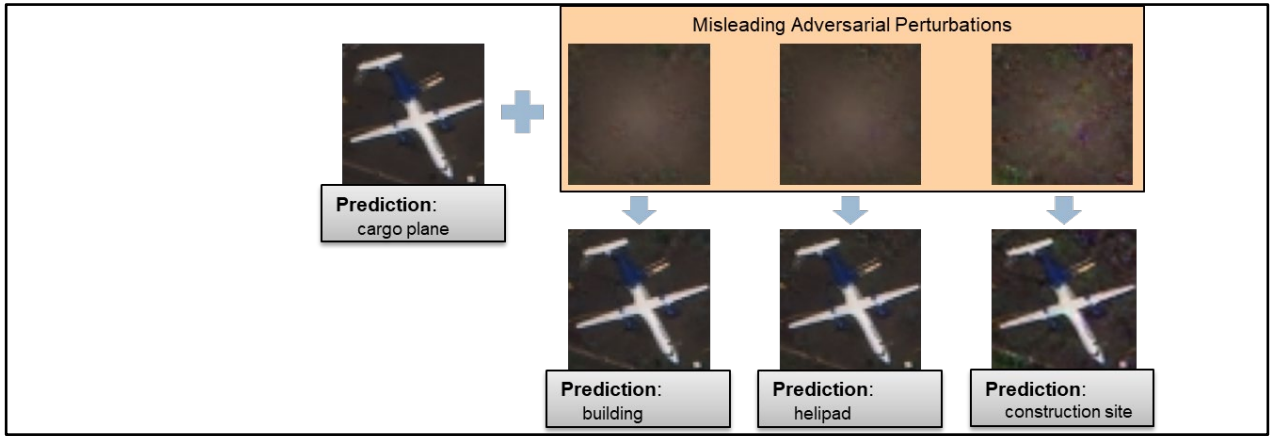


What does it mean for an AI system to be trusted and how to prove it?

Thread 2: Trusted AI Attributes and Metrics

Adversarial Robustness

- **Problem Statement:** AI/ML systems need to not only be robust to a wide variety of known inputs but must also be robust to purposefully misleading inputs.
- **Example:** AI encounters an object that is covered with material containing intentionally confusing textures that significantly affect AI prediction, such as an ATR algorithm misclassifying or not identifying a target of interest.



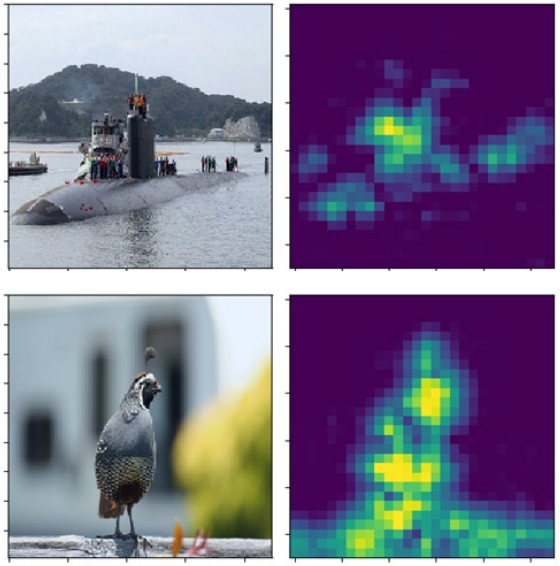
- **Description:** Adversarial robustness is the consistency of AI outputs when encountering semantically misleading data perturbations.
- **Implementation Alternatives:**
  - *IBM's Adversarial Robustness Toolbox* [11]
  - *ExamDL – AdDer (Aerospace ATIP, 2020/2021) [12]*
  - *Adversarial attacks on weather data (Aerospace, 2018)*

What does it mean for an AI system to be trusted and how to prove it?

Thread 2: Trusted AI Attributes and Metrics

Interpretability

- **Problem Statement:** AI-based systems must be instrumented in a way for users to easily understand the underlying causes of how responses were formulated.
- **Example:** A detection algorithm assigns an object as foe and initiates targeting. An easy to interpret attribution with prediction gives user confidence to allow target engagement.
- **Description:** When making a prediction or decision, interpretability is how well an AI user can understand and agree with the attribution given to an input.
  - *Users are increasingly individuals with no formal training in AI*
- **Implementation Alternatives:**
  - *Latent Representation Statistics (Aerospace*
  - *ExamDL – MEDLI (Aerospace ATIP, 2021)*
  - *Information Transfer Rate (ITR) – the agreement between a user and an algorithm, divided by the time it takes to provide a label to an input* [13]
  - *Testing for human-machine teaming with autonomous systems / HSI – Human Factors Engineering (design for usability) or performance (human-system interface)*



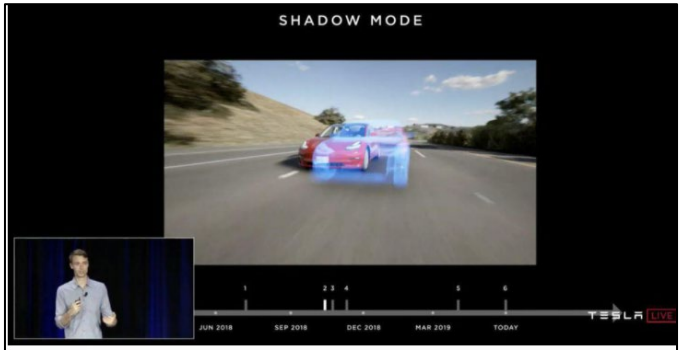
Attribution masks for the image classifier. The top image shows the input image and attribution for the correcly predicted class of 'submarine', while the lower image shows the same for the 'quail' class.

What does it mean for an AI system to be trusted and how to prove it?

Thread 2: Trusted AI Attributes and Metrics

Familiarity

- **Problem Statement:** Users must understand when to trust and when not to trust an AI/ML system. Having not enough or too much trust in an AI prediction or decision in an unsuitable environment can lead to negative consequences.
- **Example:** The Aegis Combat System runs continuously so operators can compare their decisions with system outputs, gain familiarity with scenarios that can be handled by the system, and understand when it is necessary to switch to human control
- **Description:** Familiarity is how often a user can accurately and confidently predict how an AI will operate in its deployed environment.
- **Implementation Alternatives:**
  - *Implementation of a "shadow mode" or "dark launch" for operator analysis and load testing*
    - Continuously track the degree of alignment between a user and AI predictions or actions for analysis
    - The bounds of trusted AI operation correspond to the range of potential input parameters that meets the minimum required familiarity between a user and AI
  - *Deployment and use of AI in a low-risk setting or mode prior to deployment in a higher risk environment*
  - *Evidence-Based Licensure* [4]



Tesla visualizes their "shadow mode" where the test software drives the car on a different path than it drives in reality [-] TESLA PRESENTATION

What does it mean for an AI system to be trusted and how to prove it?

Thread 2: Trusted AI Attributes and Metrics

Fairness

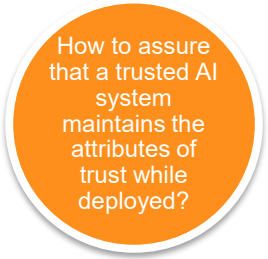
- **Problem Statement:** Deployed AI must be fair and unbiased to ensure that decisions made by the system are not unfair or do not cause unintentional negative consequences due to bias.
- **Example:** A satellite detects the presence of a nearby object originating from a foreign nation. The satellite behaves aggressively towards the object because all training data was biased towards treating foreign nation assets as hostile.
- **Description:** Fairness is the amount of bias present which may impact predictions or actions made on a population subgroup.
- **Implementation Alternatives:**
  - *Customer-funded study of data and label bias mitigation strategies for remote sensing applications (2019-2021)*
  - *Utilization of Exploratory Data Analysis (EDA) techniques on results of a ML project to quantify/prove unfairness to protected groups*
    - Microsoft's Fairlearn [14]
    - Google's Fairness Measures and Techniques for Mitigation [15]

What does it mean for an AI system to be trusted and how to prove it?

Thread 3: How to assure that a trusted AI system maintains the attributes of trust while deployed?

Monitoring

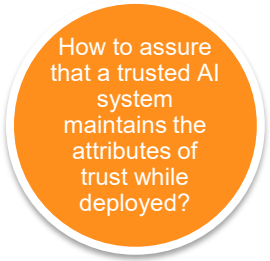
- **Problem Statement:** Over time, domain data or concepts drift from the original AI training dataset — leading to performance degradation during deployment. Systems experience a variety of failures and anomalies due to differences between the development and operational environments, interaction with other system components, and random failures.
- **Example:** A cyber security filter learns to classify between attacks and regular transient effects in a network using a training set from fall 2020. The classifier becomes less effective as tactics evolve.
- **Description:** The system must be instrumented so that data can be regularly and easily collected for AI assessment.
  - *Automated assessment of performance metrics for both proactive and reactive notifications of:*
    - AI degradation (due to model staleness or adversarial poisoning)
    - The input data changing in such a way to violate the data specification [16]
    - The AIs interaction within the operational environment has not led to an unforeseen consequence
    - Random failures within the system
- **Implementation Alternatives:**
  - *Quantify and track confidence and uncertainty for model retraining*



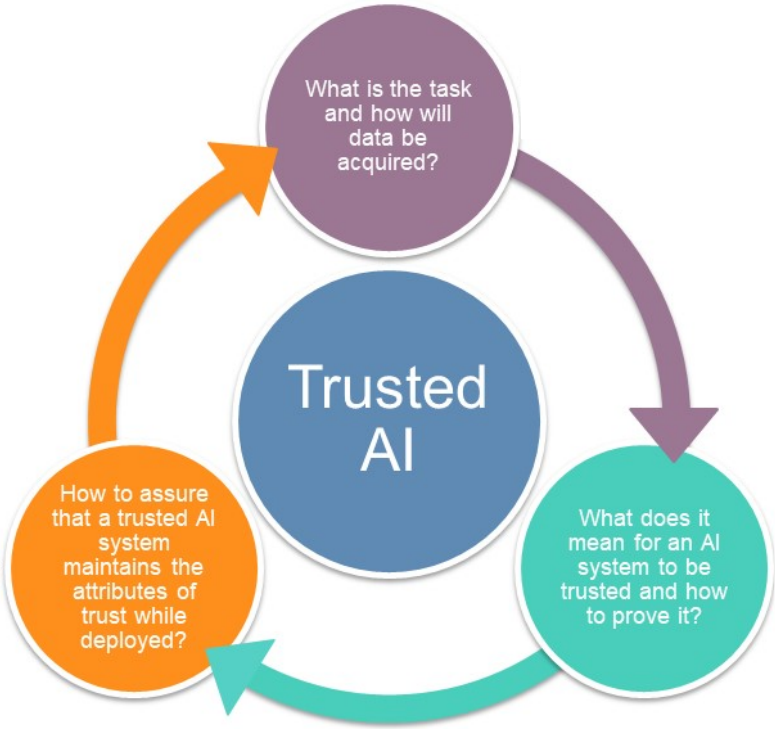
Thread 3: How to assure that a trusted AI system maintains the attributes of trust while deployed?

Control

- **Problem Statement:** When unexpected behavior occurs, some automated means of user notification and/or system interruption must be provided [17] – especially when issues arise in AI/ML that operates on rapid timelines.
- **Example:** An automated spacecraft guidance system employ a rule-based system to halt additional maneuvers when approaching nearby spacecraft.
- **Description:** Graceful termination must be defined so that interruption of the AI does not disrupt any systems relying on the AI for input.
  - *Nov 2012 OSD Directive DODD 3000.09 states that it is DOD policy that "Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force"… and "(b) Complete engagements in a timeframe consistent with commander and operator intentions and, if unable to do so, terminate engagements or seek additional human operator input before continuing the engagement*." [18]
- **Implementation Alternatives:**
  - *Deterministic backup safety-controller for autonomous systems (Aerospace ATIP, 2020/2021)*
  - *Best practices for determining control limits*
  - *Methods for test and evaluation of control methods on the system and/or architecture*
  - *Architecture-level solutions to stop failure propagation*



Applying The Threads of Trusted AI

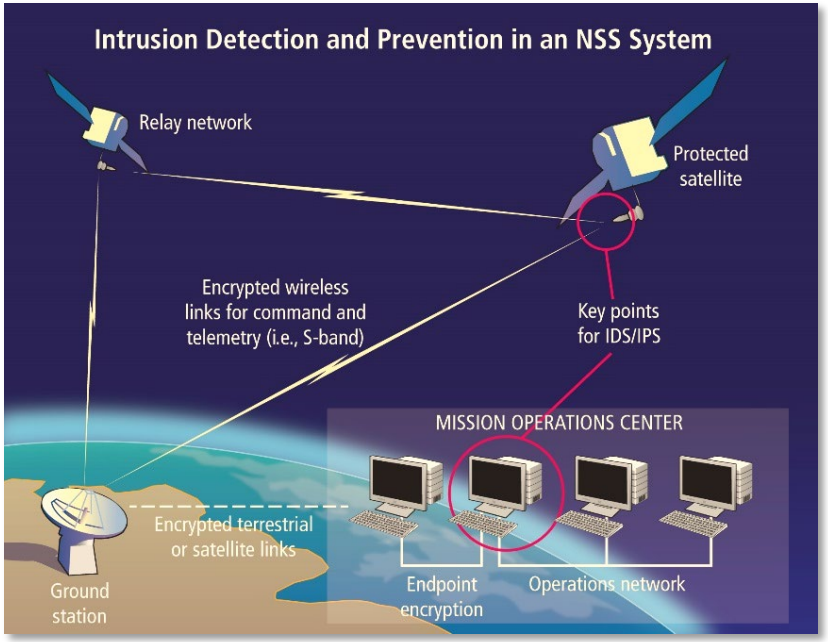


Trusted AI in Cybersecurity: An Illustration

Basic Application of the Trusted AI Framework

- **Application**: SceptreML is an Aerospace machine learning (ML)-based cybersecurity project that is in development for space ground applications
  - *The purpose is to detect anomalous information that could be indicative of cybersecurity attacks against an SV or Ground System*
  - *The ML component performs data processing and analysis to provide information to a user*

- Currently the software provides an alert to the user when anomalous activity is observed
  - *As the project advances it will also provide recommendations for actions to take based on a suite of options available within the software*
  - *Primary design consideration is whether alerts help or harm human operator efficiency*
  - *If too many false alarms need to be examined or resolved, users may end up ignoring or disabling the AI tool*



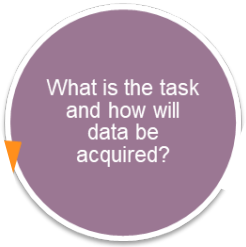
Trusted AI Framework has not yet been implemented on SceptreML, but illustrates all the threads of the Trusted AI Framework in a single context

Trusted AI in Cybersecurity: An Illustration

Thread 1: What is the task and how will data be acquired?

Objective specification:

- A satisfied objective would result in a model that provides alerts whenever network activity is not nominal:
  - Alerts should only be generated when true events of concern have occurred.
  - A poorly specified objective could result in either too many alerts swarming human operators or missing anomalous events when they occur. The cybersecurity system will then provide alerts or a selection of actions that must be chosen by a user.



- Specified objectives need to be:
  - General enough to cover a range of different operations OR
  - Be able to be adapted when network conditions change

Data specification:

- Throughout the entire lifecycle of a cybersecurity system, understanding how data were collected and used to train an AI is crucial
  - Characteristics of ground network system traffic and telemetry data will likely change over the operational lifetime
- Deliberate data collection efforts will be needed to support training an anomaly detection system on both routine and exceptional events
  - These data will also need to be updated as AI monitoring detects changes in system traffic data distributions during deployment
  - Relevant data will need to be collected to capture relevant time scales and any seasonal variations of network traffic
- Additional data should be collected when anomalous events occur
  - These would likely come from a combination of user-tagged events and labelling of data discovered by the AI
  - Addition of new data will require careful maintenance of lineage and any potential crossover between training and evaluation data

Trusted AI in Cybersecurity: An Illustration

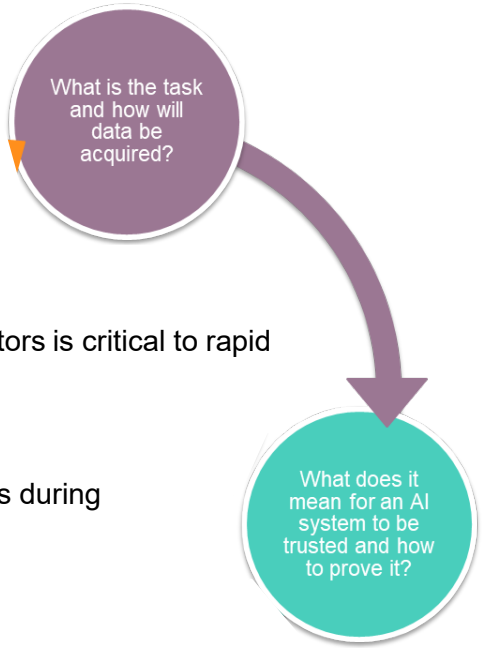
Thread 2: What does it mean for an AI system to be trusted and how to prove it?

Stability:

- The system must consistently handle the “routine” inputs that are encountered throughout normal operations. Otherwise, the cybersecurity AI may create too many alerts

Confidence and Uncertainty:

- The system must have a means to quantify the deviation from previously observed data distributions
- Additionally, thresholding could help define the boundary between routine and exceptional data, with the deviation from those thresholds defining the degree of alarm



Interpretability:

- Providing data and attribution for an anomalous event and doing so in a way that assists human operators is critical to rapid response against potential threats

Familiarity:

- Users must develop an understanding of when the system should not be heavily relied upon — such as during scheduled system maintenance that would contain approved, but atypical, traffic behavior

Adversarial Robustness:

- The detection and alerting of adversarial attacks is the primary objective of a cybersecurity system
- Damaging attacks could take the form of an injection of network traffic into the ground system that, if done in a targeted way, could gradually change the data distribution of observed traffic. Such a technique would be detrimental to the operation of a dynamic thresholding system which was used to detect anomalous events

Fairness:

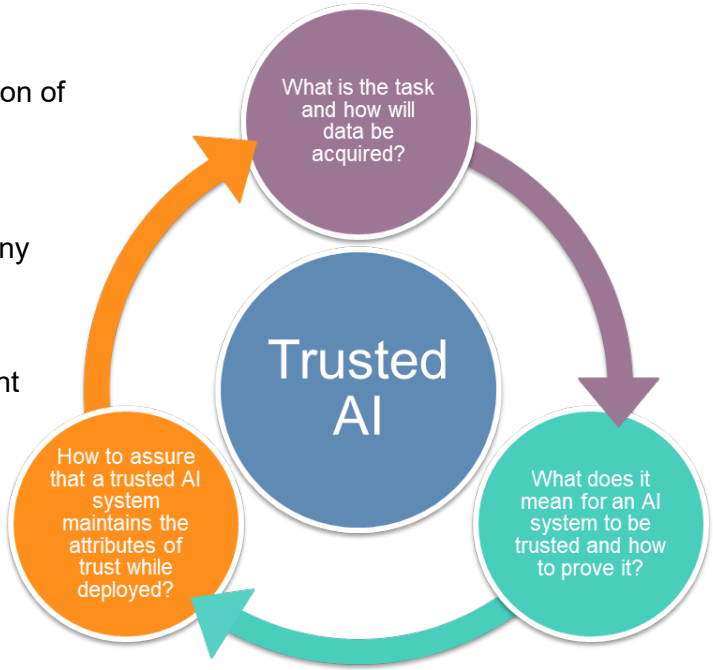
- An anomaly detection algorithm trained on past data could be strongly biased based on the limited number of anomalous events
- Any bias towards historical time periods represented in training data will lead to issues within a dynamic operational environment

Trusted AI in Cybersecurity: An Illustration

Thread 3: How to assure that a trusted AI system maintains the attributes of trust while deployed?

Monitoring

- Testing against new cyber-attack techniques are required to inform:
  - When a model needs to be retrained, or
  - If the new technique is similar enough to previous ones that the current version of the system can alert on that specific technique
- Monitoring simple metrics, such as the number of alerts, will have benefit.
  - When data shift has occurred or if the model is continually being retrained, any change in the number of alerts over time could indicate that the model has reached a sub-optimal state
- Regular retraining or having different anomaly detection systems in place for different tasks could mitigate the issue of task-dependent network conditions

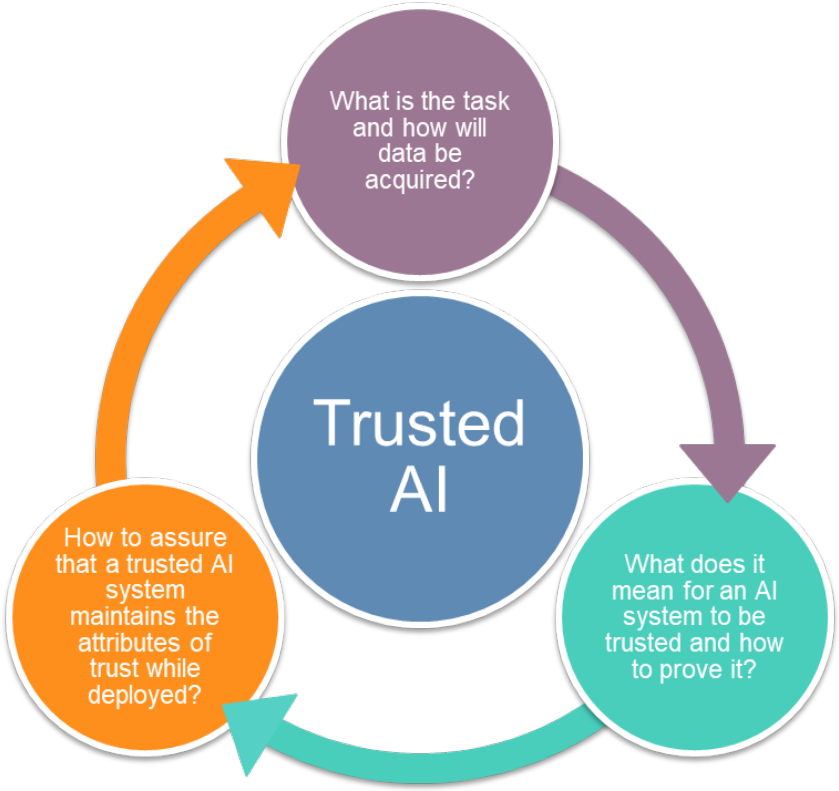


Control

- All cybersecurity systems operate within a larger environment – users must able to intervene and terminate some security systems gracefully
- Turning off a system that only provides alerts should have minimal impact on a network, but any downstream consumers of alerts would need to be considered.

Conclusion

- Trusted AI Framework written for general applicability
- Some applications emphasize components of trust not explicitly discussed here:
  - *Some applications necessitate a strong emphasis on security and privacy.*
  - *Others will require frequent cooperation with users, requiring a deeper focus on human-machine teaming (e.g., chatbots and robotic assistants)*
- Generating an AI/ML software strategy should complement a broader program strategy
  - *This includes proper data strategy and verification and validation methods*
- As AI is more widely deployed, concerns of managing performance expectations will continue to increase
  - *The larger architecture must be resilient enough to avoid failure in the event of failure of an individual AI/ML agent or agent-based system*



We offer a framework as a starting point for creating procedures to generate, test, and monitor systems that use AI/ML in order to better trust them

Backup

Definitions

- **Adversarial Robustness** — the AI's ability to provide outputs consistent with those generated when no deceptive perturbations are present along with the ability to detect when such perturbations are present.
- **Artificial Intelligence (AI)** — the subdiscipline of computer science focused of the development of hardware and software-based solutions which are capable of successfully performing tasks typically associated with human-level cognition or intelligence.
- **Confidence and Uncertainty** — quantification of model sureness across the input space along with the ability to discern when inputs fall outside of the typical data distribution.
- **Fairness** — not providing favorable or unfavorable outcomes to only a subset of represented data.
- **Familiarity** — a user's ability to anticipate the predictions or decisions an AI-based application will provide.
- **Interpretability** — the degree to which a user can understand the cause of an AI algorithm prediction.
- **Machine Learning (ML)** — a branch of artificial intelligence focused on building models from data for purposes such as pattern recognition, prediction, capturing latent structure, or defining action policies.
- **Stability** — is the consistency of model performance when provided inputs that fall within a routine range of data parameters.

References

1. Minsky, Marvin. *The Emotion Machine Common Sense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, 2007.
2. Naughton, John. “To err is human – is that why we fear machines that can be made to err less?.”14 December 2019 https://www.theguardian.com/commentisfree/2019/dec/14/err-is-human-why-fear-machines-made-to-err-less-algorithmic-bias (Accessed 7 March 2021).
3. DeVries, Byron, and Betty HC Cheng. "Automatic detection of incomplete requirements via symbolic analysis." Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems. ACM, 2016.
4. Tate, David M., et al. A Framework for Evidence-Based Licensure of Adaptive Autonomous Systems: Technical Areas. Institute for Defense Analyses Alexandria, 2016.
5. Friedler, Sorelle A., et al. "Assessing the Local Interpretability of Machine Learning Models." arXiv preprint arXiv:1902.03501 (2019).
6. Leike, Jan, et al. "Scalable agent alignment via reward modeling: a research direction." arXiv preprint arXiv:1811.07871 (2018).
7. Gebru, Timnit, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2020).
8. Castro, Daniel C., Ian Walker, and Ben Glocker. "Causality matters in medical imaging." Nat Commun 11, 3673 (2020).
9. Aerospace and Mitre document, Aerospace TOR-2020-0180. (Unpublished)
10. Gal, Y. (2015, June 6). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ArXiv.Org. https://arxiv.org/abs/1506.02142
11. Nicolae, Maria-Irina, et al. “Adversarial Robustness Toolbox v1. 0.0.” arXiv preprint arXiv:1807.01069 (2018). (IBM’s Adversarial Robustness Toolbox https://adversarial-robustness-toolbox.readthedocs.io/en/stable/)
12. Wendoloski, Eric B. ExamDL/AdDer: Practical Guide for Utilizing Explainable Methods for Deep Learning, The Aerospace Corporation. ATR-2020-01500. September 2020. (limited distribution)
13. Moraffah, Raha, et al. "Causal Interpretability for Machine Learning-Problems, Methods and Evaluation." ACM SIGKDD Explorations Newsletter 22.1 (2020): 18-33.
14. Microsoft’s Fairlearn Toolbox https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
15. Google AI Practices https://ai.google/responsibilities/responsible-ai-practices/
16. Suprem, Abhijit. "Concept Drift Detection and Adaptation with Weak Supervision on Streaming Unlabeled Data." arXiv preprint arXiv:1910.01064 (2019).
17. Danks, David, and Alex John London. "Regulating autonomous systems: Beyond standards." IEEE Intelligent Systems 32.1 (2017): 88-91.
18. DOD Directive 3000.09 “Autonomy in Weapon Systems,” November 12, 2012. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

AEROSPACE REPORT NO.
ATR-2021-01456

A Framework for Trusted Artificial Intelligence in High-Consequence Environments

Cognizant Program Manager Approval:

Brian E. Hardt, GENERAL MANAGER
ENGINEERING & TECHNOLOGY GROUP
OFFICE OF EVP

Aerospace Corporate Officer Approval:

Todd M. Nygren, SENIOR VP ENGINEERING & TECHNOLOGY
OFFICE OF EVP

SI0669

AEROSPACE REPORT NO.
ATR-2021-01456

A Framework for Trusted Artificial Intelligence in High-Consequence Environments

Content Concurrence Provided Electronically by:

Lauren H. Perry, SENIOR PROJECT ENGINEER
SPACE APPLICATIONS
SURVIVABILITY & RESILIENCE
NATIONAL SYSTEMS GROUP

Office of General Counsel Approval Granted Electronically by:

Kien T. Le, ASSISTANT GENERAL COUNSEL
OFFICE OF THE GENERAL COUNSEL
OFFICE OF GENERAL COUNSEL & SECRETARY

SI0669

AEROSPACE REPORT NO.
ATR-2021-01456

A Framework for Trusted Artificial Intelligence in High-Consequence Environments

Export Control Office Approval Granted Electronically by:

Angela M. Farmer, SECURITY SUPERVISOR
GOVERNMENT SECURITY
SECURITY OPERATIONS
OFFICE OF THE CHIEF INFORMATION OFFICER

SI0669

All comments will be made public as-is, with no edits or redactions. Please be careful to not include confidential business or personal information, otherwise sensitive or protected information, or any information you do not wish to be posted.

Comment Template for Responses to NIST Artifical Intelligence Risk Management Framework

Submit comments by August 19, 2021:

General RFI Topics (Use as many lines as you like)	**Response #**	**Responding organization**	**Responder's name**	**Paper Section (if applicable)**	**Response/Comment (Include rationale)**	**Suggested change**
Aerospace FFRDC Response to RFI	**1**	**The Aerospace Corporation (FFRDC)**	**Mr. James Myers**	**Applicable to all - best placed as addendums**	**Aerospace presents two pulicly released documents - a publication and a slide presentation**	**Please see Cover Letter and two attachments**
Responses to Specific Request for information (pages 11,12, 13 and 14 of the RFI)						
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;						

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;						
3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: transparency, fairness, and accountability;						
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;						

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;						
6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;						
7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;						
8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation – and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.						

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, “AI RMF Development and Attributes”);						
10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include – but are not limited to – the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and						
11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.						

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.						