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

Distribution Statement A: Approved for public release; distribution unlimited

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?

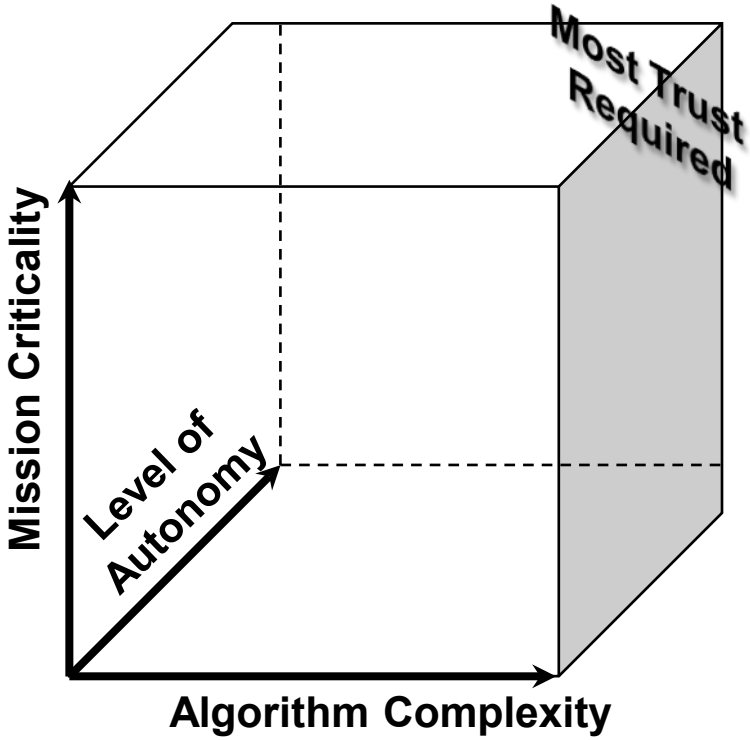
Why Trust Matters in Safety Critical Systems

- 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

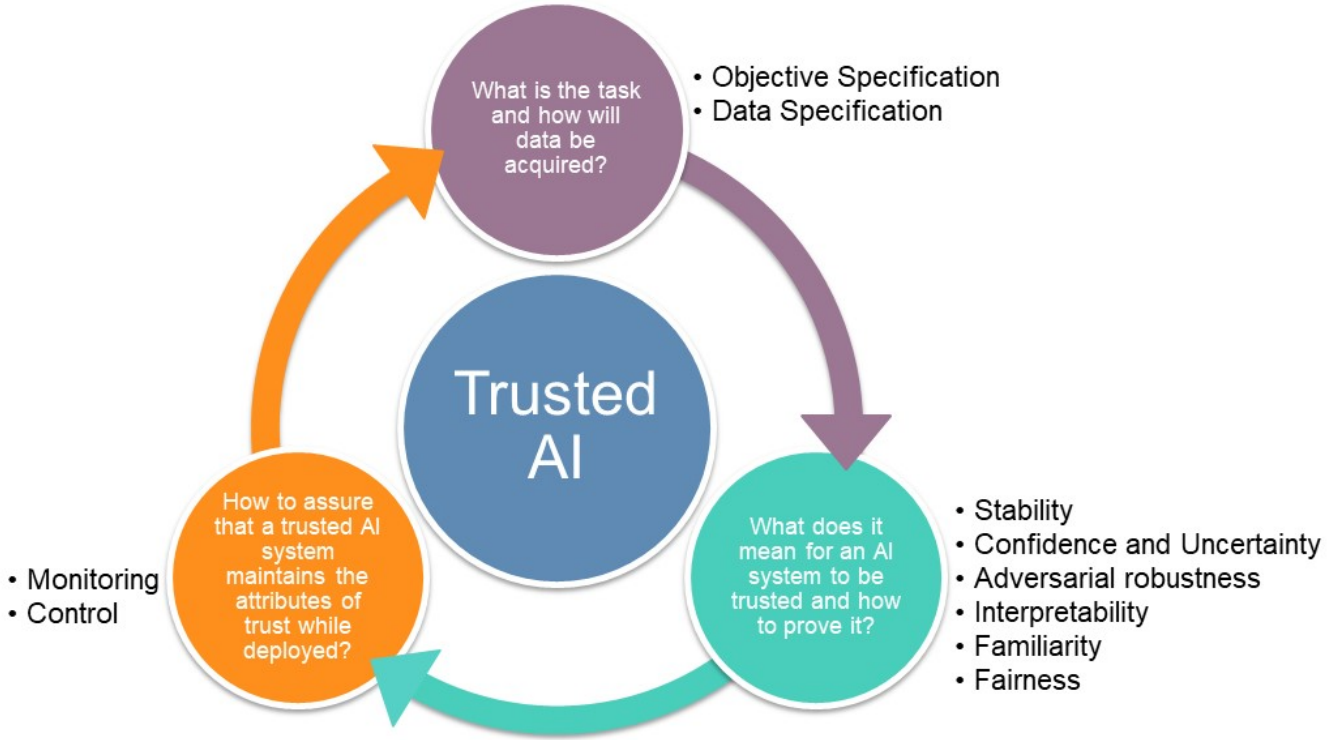
- Until recently, most research has focused almost exclusively on Adversarial and Explainable AI. What changed?
  - *Enforcement of General Data Protection Regulation (GDPR) in the EU since 2018 requires not only "right to explanation" from algorithmic decisions, but also prohibits processing data that 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)

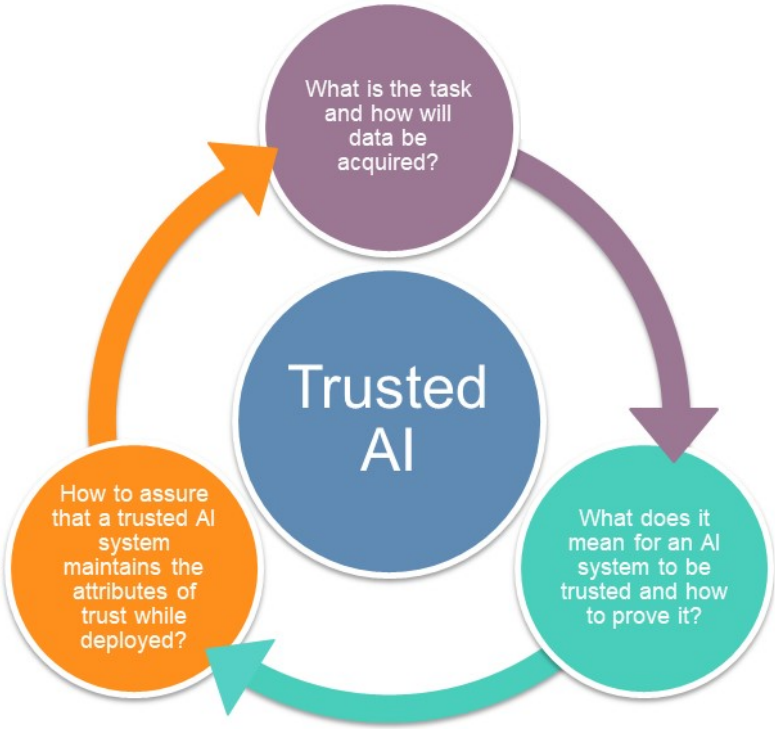
Trusted AI is as much a philosophy and engineering process as it is a system feature

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)*

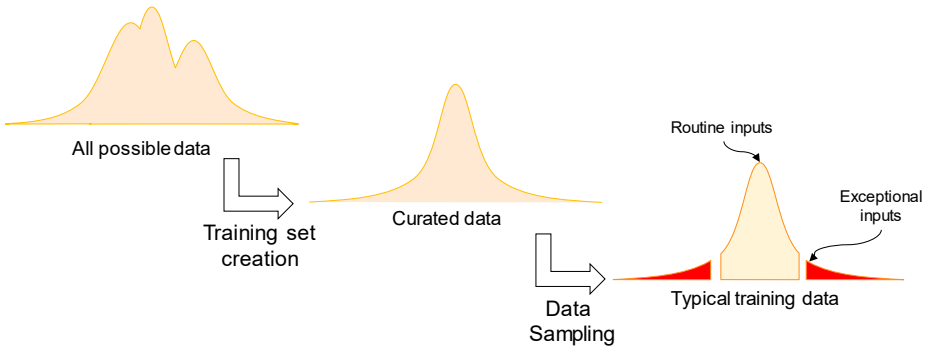
Objective Specification will provide the groundwork for defining the standard by which the AI will be assessed

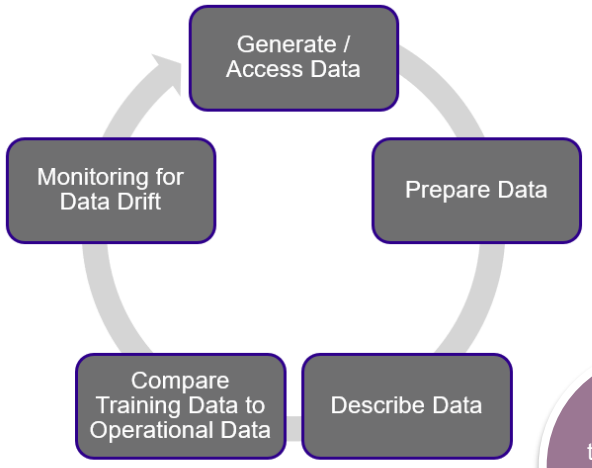
What is the task and how will data be acquired?

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

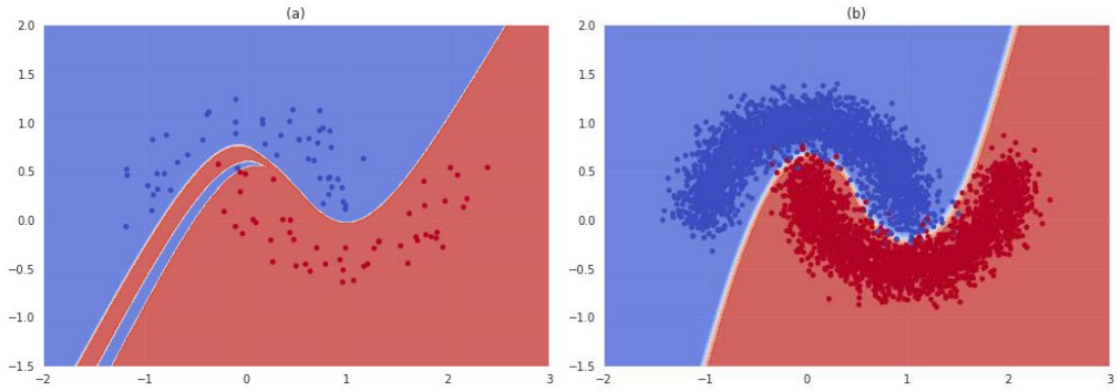
What is the task and how will data be acquired?

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



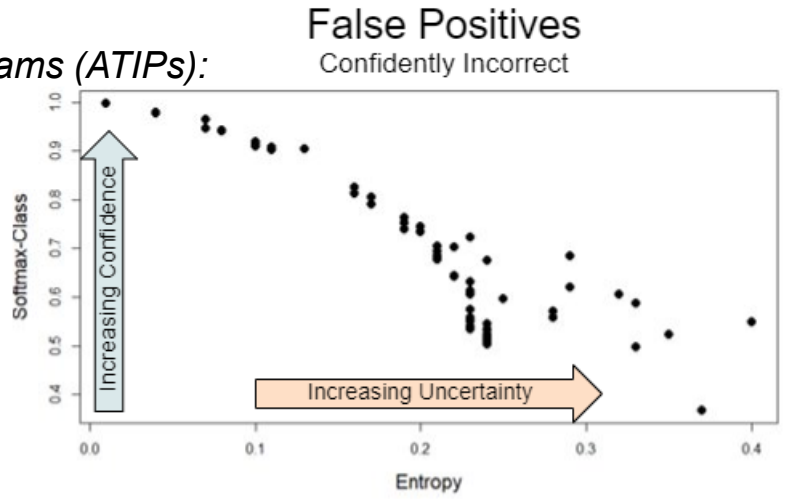
Training data and decision boundaries from two training runs using different sample sizes. (a) Sample size of 100 resulted in overly complex decision boundary. (b) Sample size of 5000 resulted in simpler, but more accurate decision boundary.

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)

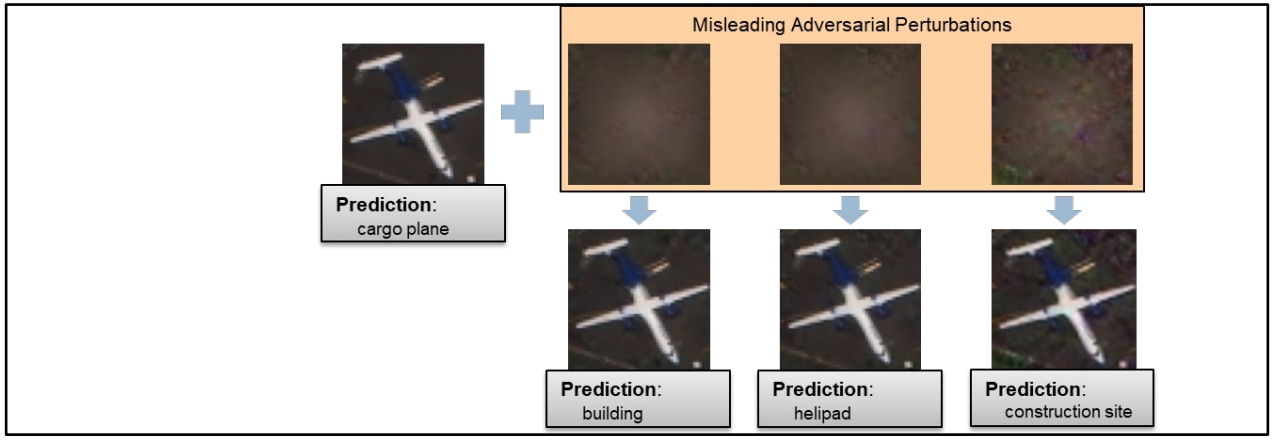


What does it mean for an AI system to be trusted and how to prove it?

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.



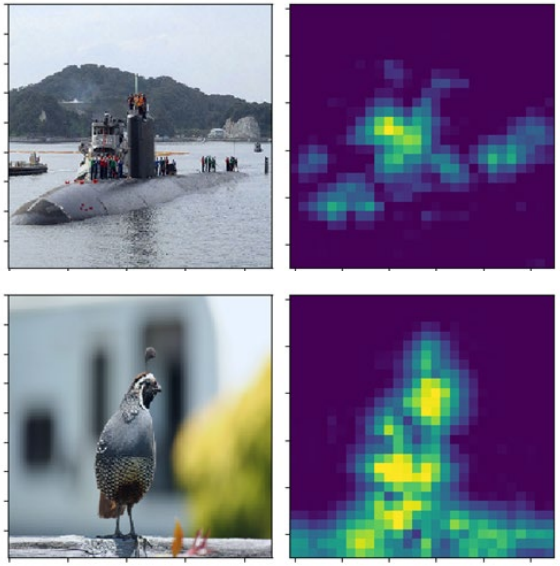
- **Description:** Adversarial robustness is the consistency of AI outputs when encountering semantically misleading data perturbations.
- **Implementation Alternatives:**
  - *IBM's Adversarial Robustness Toolbox* [11]
  - *ExamDL – AdDer (Aerospace ATIP, 2020/2021) [12]*
  - *Adversarial attacks on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



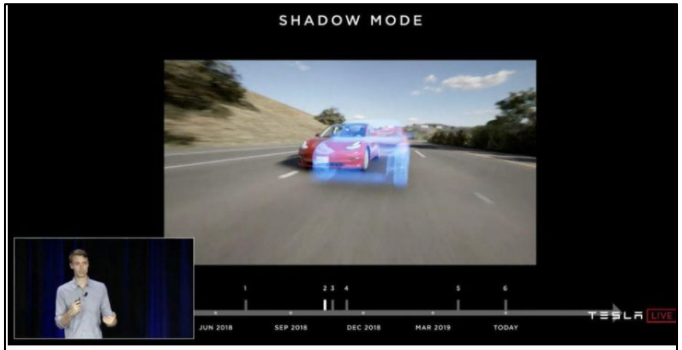
Attribution masks for the image classifier. The top image shows the input image and attribution for the correcly predicted class of 'submarine', while the lower image shows the same for the 'quail' class.

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

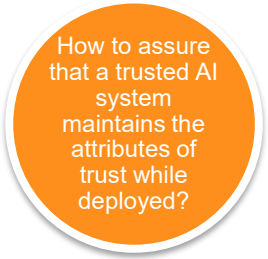
- **Problem Statement:** Deployed AI must be fair and unbiased to ensure that decisions made by the system are not unfair or do not cause unintentional negative consequences due to bias.
- **Example:** A satellite detects the presence of a nearby object originating from a foreign nation. The satellite 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

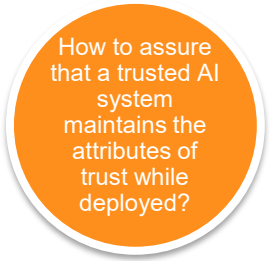
- **Problem Statement:** Over time, domain data or concepts drift from the original AI training dataset — leading to performance degradation during deployment. Systems experience a variety of failures and anomalies due to differences between the development and operational environments, interaction with 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*



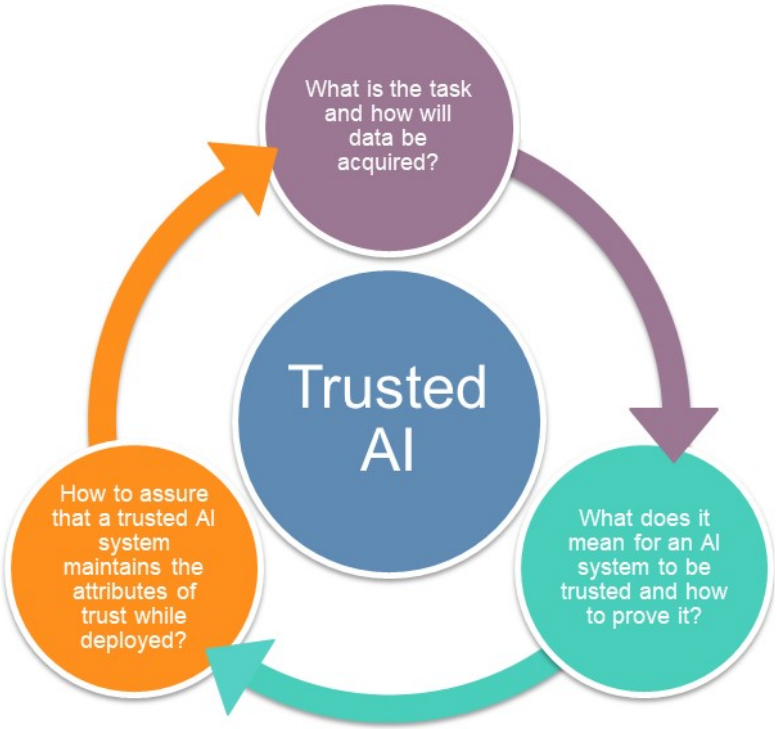
Thread 3: How to assure that a trusted AI system maintains the attributes of trust while deployed?

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

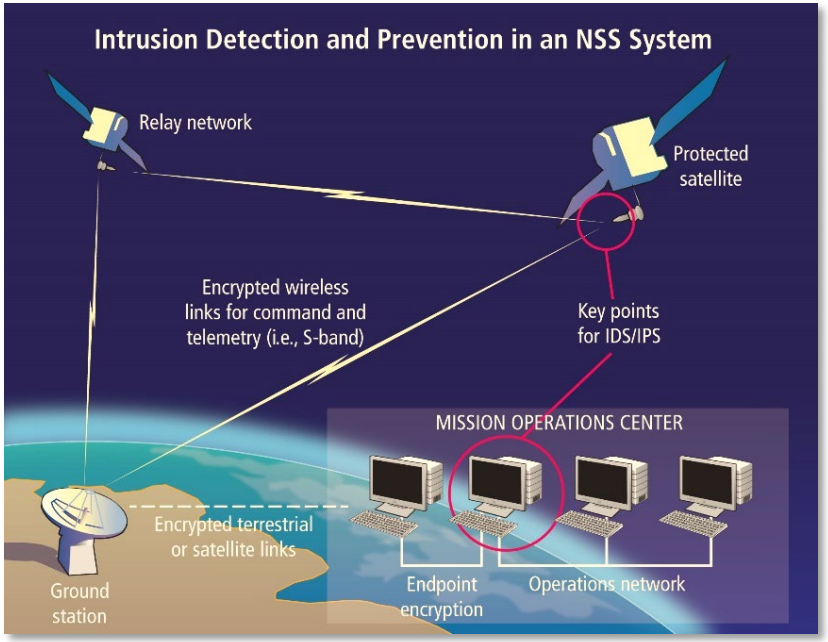


Trusted AI in Cybersecurity: An Illustration

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*



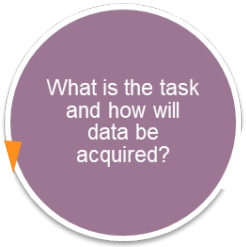
Trusted AI Framework has not yet been implemented on SceptreML, but illustrates all the threads of the Trusted AI Framework in a single context

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

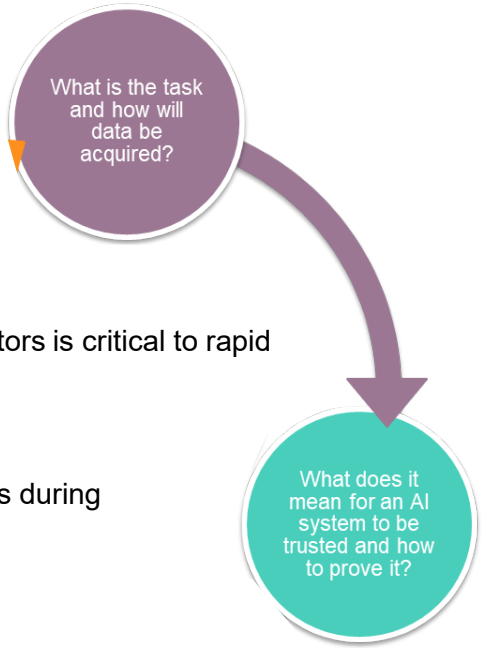
Thread 2: What does it mean for an AI system to be trusted and how to prove it?

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:

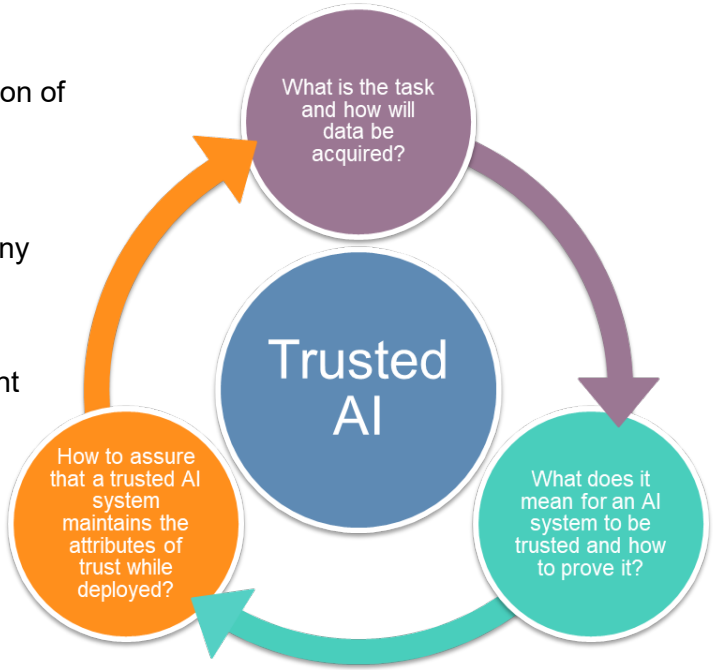
- An anomaly detection algorithm trained on past data could be strongly biased based on the limited number of anomalous events
- Any bias towards historical time periods represented in training data will lead to issues within a dynamic operational environment

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

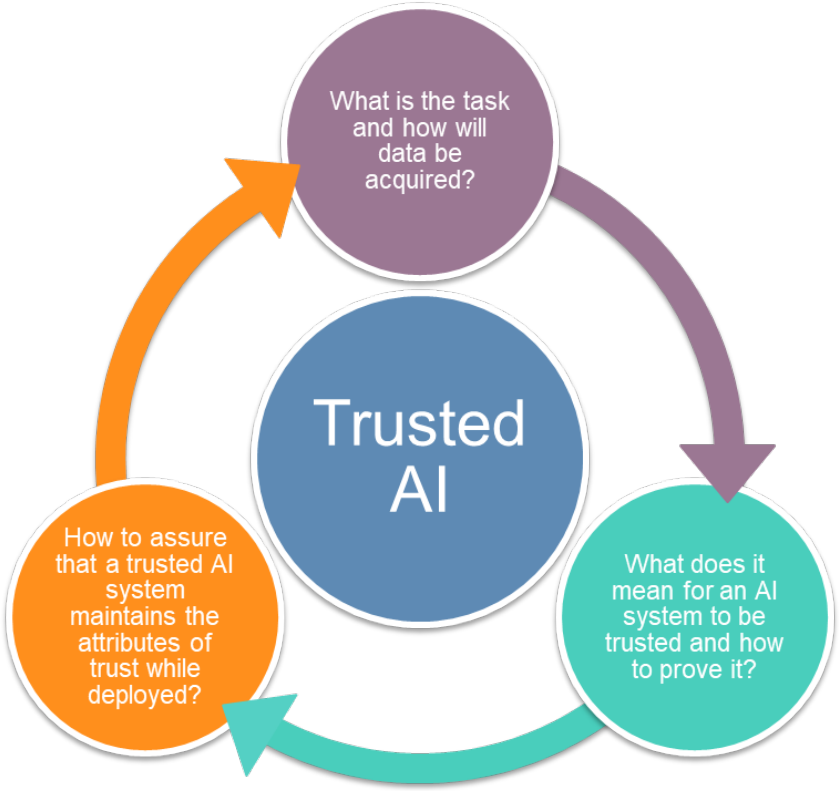


Control

- 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

© The Aerospace Corporation, 2021.

All trademarks, service marks, and trade names are the property of their respective owners.

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

© The Aerospace Corporation, 2021.

All trademarks, service marks, and trade names are the property of their respective owners.

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

© The Aerospace Corporation, 2021.

All trademarks, service marks, and trade names are the property of their respective owners.

SI0669