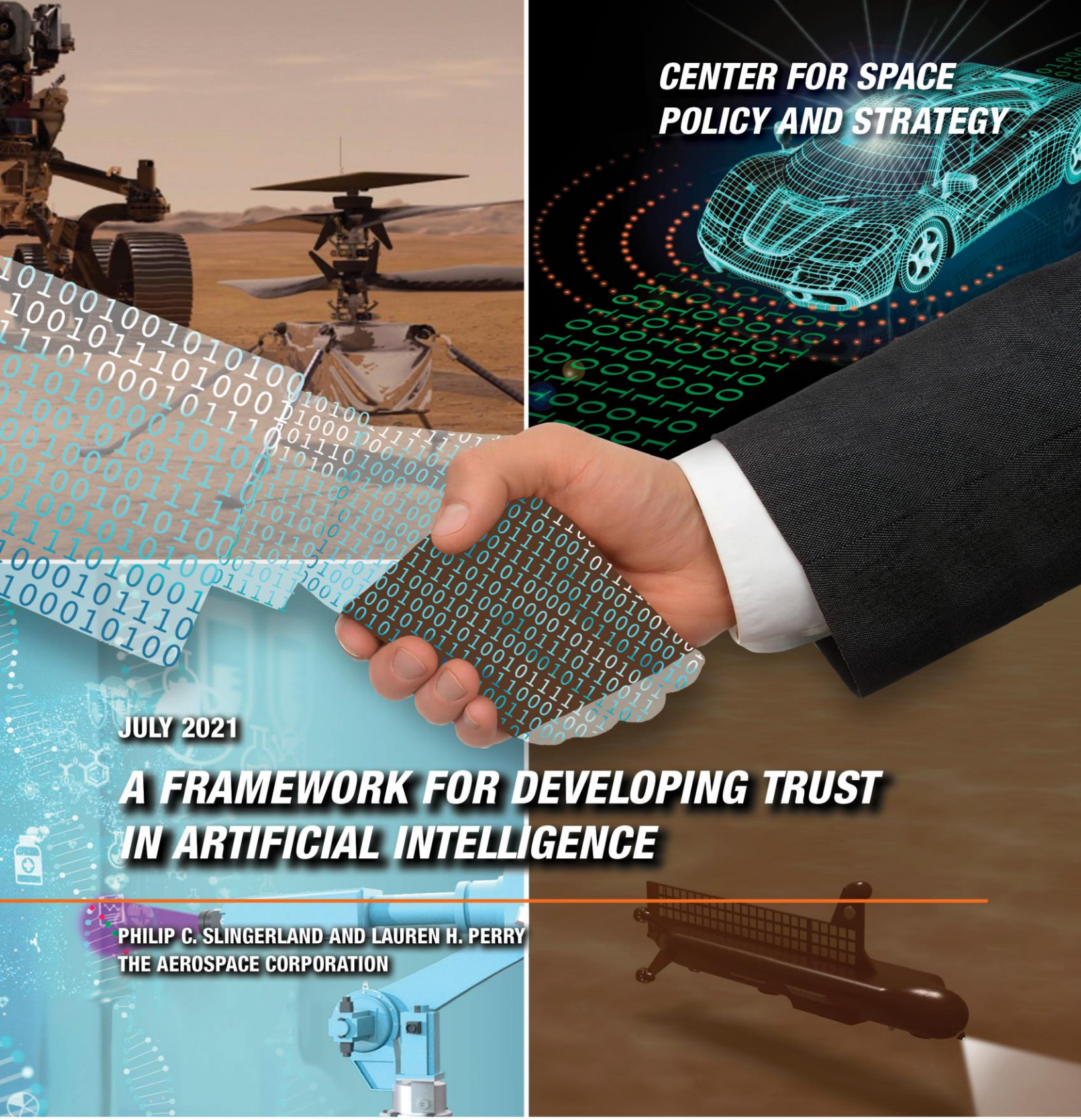

JULY 2021

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

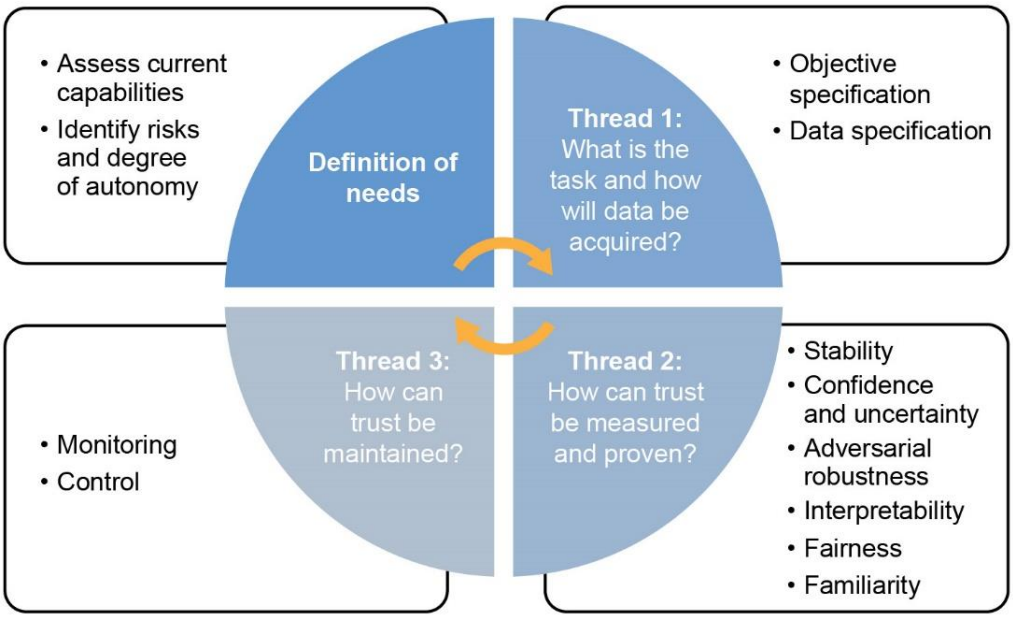

Figure 1: The threads of Trusted AI as they are applied within the development lifecycle. As AI is deployed in real-world environments, this cycle can repeat as new needs or AI performance limitations emerge.

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.