USING STANDARDS TO MITIGATE RISKS

2018
PUBLIC-PRIVATE
ANALYTIC EXCHANGE PROGRAM

contents

KEY INSIGHTS 3

INTRODUCTION TO AI & NATIONAL SECURITY 4

RESEARCH APPROACH 7

FRAMEWORK FOR AI RISK 8

THREE ADVERSARIAL SCENARIOS 9

- IMAGE RECOGNITION 10
- NATURAL LANGUAGE PROCESSING 11
- PATTERN RECOGNITION 12

VULNERABILITIES & RISKS 13

STANDARDS TO MITIGATE RISKS 14

A PLACE TO START 16

OUTLOOK: A DIALOGUE 18

WAY FORWARD: DECISION GUIDE ON AI STANDARDS 19

TEAM MEMBERS 20

KEY INSIGHTS

This study seeks to identify factors to consider when formulating standards to manage the national security risks of using artificial intelligence (AI) for decision support systems. The goal is to start a dialogue on creating standards that will reduce the risk from use, misuse, and exploitation of AI, without impeding the United States' technological development and competitive advantage. Based on our literature review and interviews with key stakeholders, we focused our report on four factors.

Standards. Potential standards should: 1) be flexible to keep pace with innovation; 2) focus on areas with large scale usage that address specific risks and impacts; 3) strike a balance between standardized testing and rapid iterative development with users and; 4) create buy-in within all appropriate private and public sector entities.

People. Public/private partnerships and education for users is critical to understanding AI technology and the associated threats, vulnerabilities, and risks. Important activities should include creating best practices while mitigating political obstacles and establishing common language to talk about the same subjects or issues in the same way. For AI weapons systems, the focus should include finding better models for what a machine/human partnership looks like and determining where best to keep the human in the loop.

Data. The backbone for most AI systems depends on the quantity and quality of data. However, the proprietary nature of the data and software often hinders sharing. Among the steps to address data issues include identifying and creating an incentive structure for cooperation; building a flexible and open data ecosystem with standards for structuring and labeling data; focusing more on data collection and annotation processes than end-products; and finding a balance between privacy protection, data regulations, safeguards, and data-driven research.

Algorithms. To build trust, fairness, transparency, and accountability of AI to curtail error and misuse while ensuring functionality and securing against attacks, algorithms should undergo the "illities" test. The test looks at reliability, accountability maintainability, functionality, debug-ability, evolve-ability, fragility, and vulnerability. Algorithms should also incorporate ethical, legal, privacy, transparency, and bias concerns. The main challenge is the complexity and difficulty in getting the public and private sector that have different incentive structure to agree on specific standards.

INTRODUCTION

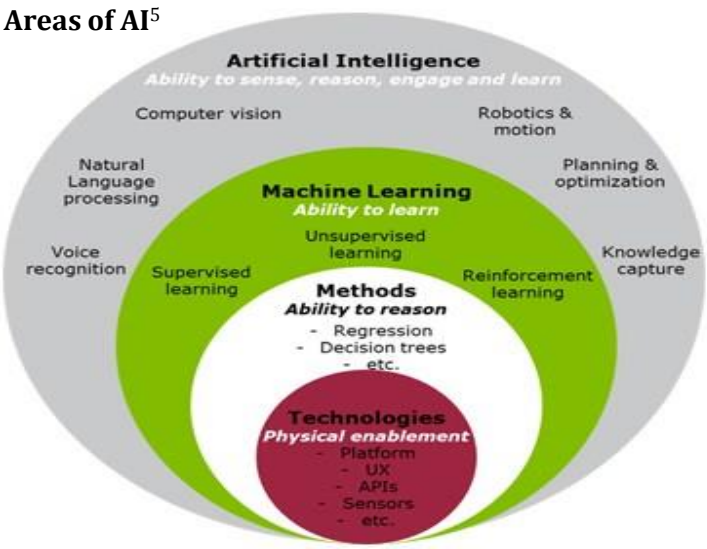
Artificial intelligence (AI) likely is one of the most dramatic technological game changers of our time with the potential to transform human life from daily social interactions to how we conduct warfare. Specifically, AI will play a critical role in driving change in military, information, economic superiority, and the nature of security risks that will affect the ability of the United States to pursue its four pillars of national security.

- Progress in AI has the potential to transform the United State's economy and security, including employment, education, public safety, and national security, possibly dictating new strategy, organization, priorities, and allocated resources.[1] The implications of adversaries' abilities to use AI are potentially broad and profound as they can "more readily develop weapon systems that can strike farther, faster, and harder and challenge the United States in all warfare domains."[2]

- AI will have digital, physical, and political security implications, expanding existing threats, introducing new threats, and changing the character of threats and of war. These changes could include the automation of social engineering attacks, vulnerability discovery, influence campaigns, terrorist repurposing of commercial AI systems, increased scale of attacks, and manipulation of information availability.[3]

AI Definition: devices and systems that have some kind of ability to plan, reason and learn, sense and build some kind of perception of knowledge and communicate in natural language. For this study, AI can be devices or decision-making aids that rely on automated, data-driven, or algorithmic learning procedures.[4] See below graphic for different aspect of AI.

Figure 1. Different Areas of AI[5]



[1]Greg Allen and Taniel Chan, *Artificial Intelligence and National Security*, A study on behalf of Dr. Jason Matheny, Director of the U.S. Intelligence Advanced Research Projects Activity (IARPA), July 2017
[2]Statement for the Record, *Worldwide Threat Assessment of the US Intelligence Community*, February 13, 2018.
[3]For further discussion of the changing threats see "*The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*," February 2018
[4]Osonde A. Osoba, William Welser IV, *The Risks of Artificial Intelligence to Security and the Future of Work*, RAND, 2017
[5] Stefan van Duin & Naser Bakhshi, *Part 1: Artificial Intelligence Defined*, Deloitte, 28 March 2017

EXPLOSIVE GROWTH OF AN INDUSTRY

Use of AI is growing rapidly all over the world. According to a forecast by Tractica, LLC, a market intelligence company, the revenue generated from the direct and indirect application of AI software is estimated to grow from $643.7 million in 2016 to $36.8 billion by 2025 (see Figure 2). This represents a significant growth curve for the 9-year period with a compound annual growth rate (CAGR) of 56.8%.[1] Figure 3 shows where 100 of the most promising startups, as determined by CB Insights, are using AI to transform industries.[2]

Figure 2.

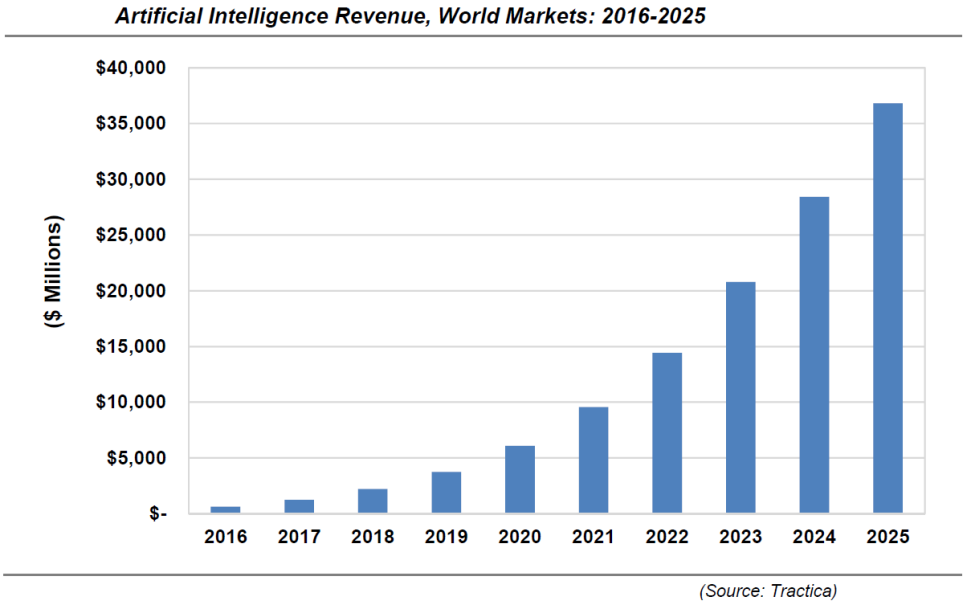
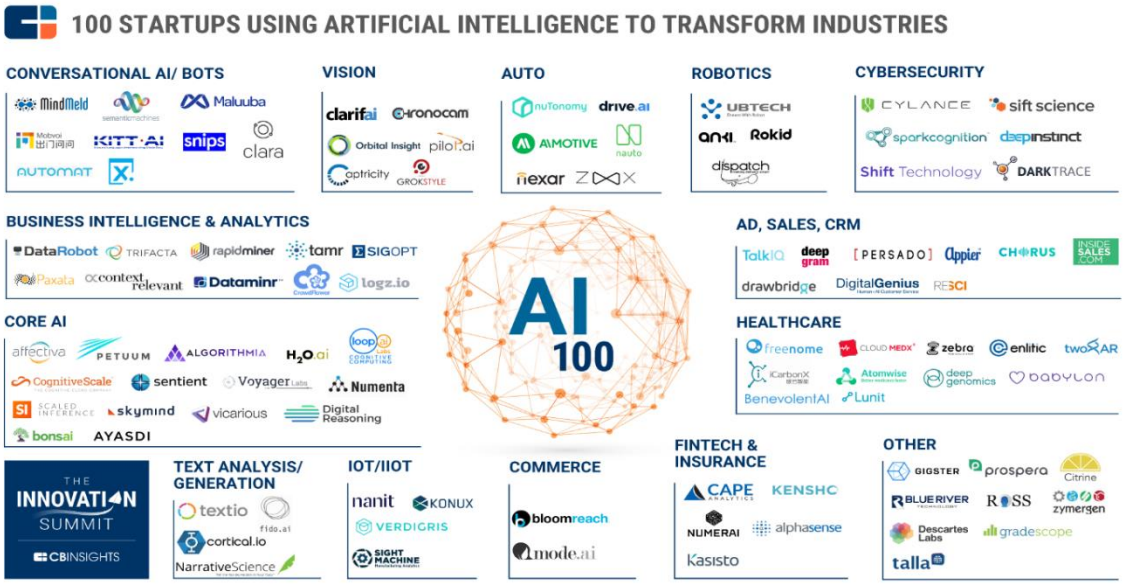


Figure 3.



[1] Tractia, https://www.tractica.com/wp-content/uploads/2016/08/MD-AIMF-3Q16-Executive-Summary.pdf
[2] CB Insights, https://www.cbinsights.com/research/artificial-intelligence-top-startups/

NATIONAL SECURITY IMPLICATIONS

National security interests traditionally fall within the categories of defense of the homeland, economic well-being, favorable world order, and promotion of values, which are articulated in the 2017 National Security Strategy (NSS), the 2018 National Defense Strategy (NDS), and the 2014 Quadrennial Homeland Security Review. In this paper, security implications refer to the impact of AI in these national security areas.

- The NSS establishes four pillars of national security: 1) Protecting the American People, the Homeland, and the American Way of Life; 2) Promoting American Prosperity; 3) Preserving Peace Through Strength and; 4) Advancing American Influence.[1]

- Among the key NDS objectives include defending the homeland; dissuading, preventing, or deterring state adversaries and non-state actors from acquiring, proliferating, or using weapons of mass destruction; and preventing terrorists from directing or supporting external operations against the United States homeland and our citizens, allies, and partners overseas.[2]

- The 2014 Quadrennial Homeland Security Review identifies five missions for Homeland Security: 1) Preventing Terrorism and Enhancing Security; 2) Managing U.S. Borders; 3) Administering Immigration Laws; 4) Safeguard and Secure Cyberspace; and 5) Strengthening National Preparedness and Resilience.[3]

[1] *National Security Strategy of the United States of America*, https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf, 2017
[2] *National Defense Strategy of the United States of America*, https://www.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf, 2018
[3] *The 2014 Quadrennial Homeland Security Review*, https://www.dhs.gov/sites/default/files/publications/2014-qhsr-final-508.pdf, 2014

RESEARCH APPROACH

To understand governance, development, technology, algorithms, and potential standards, we conducted a literature review and interviewed stakeholders, including practitioners, developers, researchers, academics, and end-users from 19 organizations.

- Narrowing the Focus. We conducted a review of literature on AI to understand the existing work done in the field, scope our project and ensure that we did not duplicate existing analysis. "Customer personas" helped us guide and frame our research to be of greatest utility to a potential reader. *These exercises led us to focus on AI used for decision support because of its prevalence in the defense sector.* AI decision support applications help people make effective decision based on their circumstances. These systems do so by learning based on a person's inputs while the person's decision becomes more accurate based on the information the system provides.[1]

- Identification of experts. We then identified specific policymakers, practitioners, developers, researchers, academics, and end-users who could help us better understand governance, development, technology, algorithms, and potential standards.

- Planning. We collaboratively developed a set of questions to frame our discussions with AI subject matter experts. These focused on understanding the technology and their thoughts on standards.

- Implementation. We conducted interviews with individuals from seven public sector organizations and twelve in the private sector and academia. These individuals represented a wide range of interests and businesses including Commerce, Consulting, Defense, Hardware, Healthcare, Homeland & National Security, Information Security, Policy, Standards, Technology Development, and Vision.

ACKNOWLEDGEMENTS: We are very grateful for all the unique insight we received from interviewees who contributed to this report and we take full responsibility for all errors of fact or interpretation implied or explicit in this paper.

[1]MIT Technology Review Insights, *AI at a Glance*, https://www.technologyreview.com/profile/mit-technology-review-insights/, June 20, 2016

FRAMEWORK FOR AI RISK

We developed a risk framework to map the relationships among ideas relating to risk and provide a parallel example related to AI. This framework provides a picture of where potential threats, vulnerabilities, and risks can occur, enabling a better understanding of the security impacts and where to possibly apply standards. We view **risk** as a function of the **likelihood** of a given **threat-source** exercising a potential **vulnerability**, and the resulting **impact** of that adverse event on security.[1]

- The outer ring of Figure 4 represents a cycle of generalized risk. Starting with Assets that are compromised by Threats, which exploit Vulnerabilities that are exposed to Risks, which are mitigated by Controls to protect Assets.

- The inner ring represents the same cycle for an example AI-related system. Start with Training Data compromised by Actors who feed malicious data to AI systems, which are Designed with narrow purposes and algorithms that lack pressure testing. Decisions derived from algorithms that are not scrutable require Standardized architecture to protect the compromised Training Data.

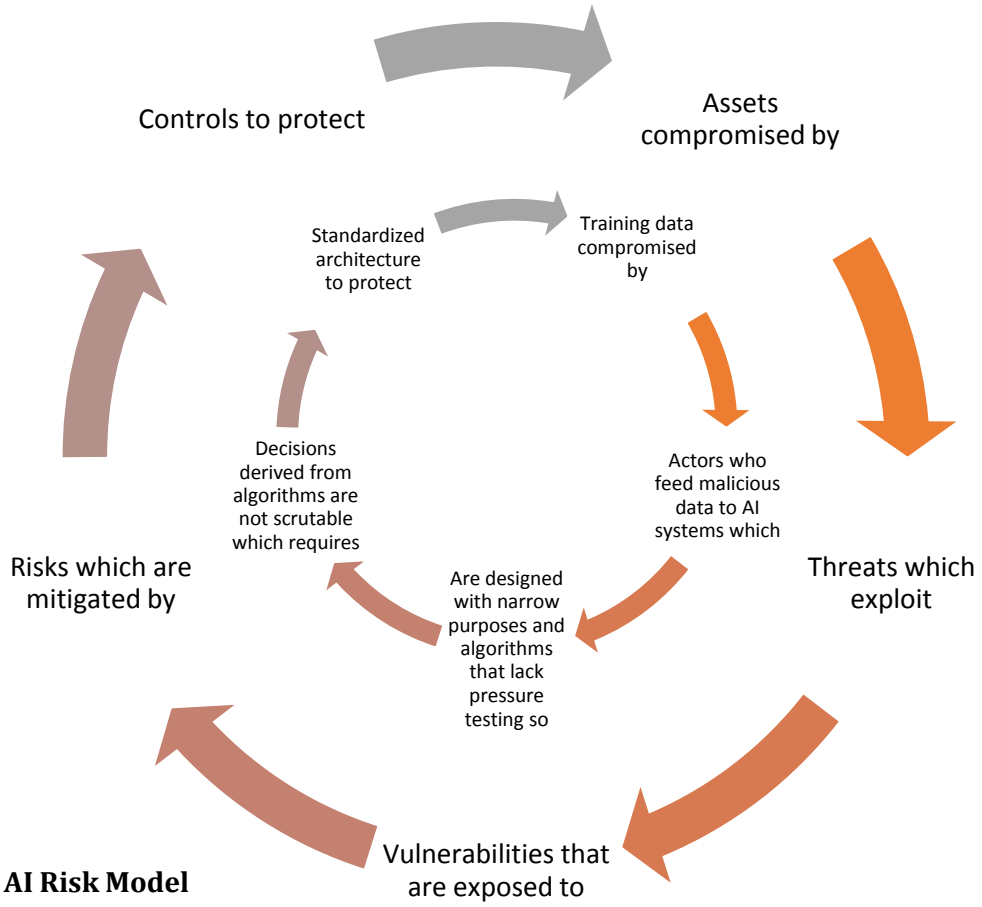


Figure 4. AI Risk Model

[1]Joint Task Force Transformation Initiative, *Guide for Conducting Risk Assessments*, National Institute of Standards and Technology, 2012

THREE ADVERSARIAL SCENARIOS

We created three scenarios to illustrate how adversaries can potentially target AI decision support system's vulnerabilities within the defense sector's image recognition, anomaly detection, and natural language processing capabilities and create security risks. Applying AI-enabled tools for national security purposes without understanding the potential threats, vulnerabilities, and risks associated with these models can lead to destructive consequences in the future.

Threat actors can highjack data sets or training processes and manipulate them to yield advantageous results once the tool is implemented in the wild. Maladaptation of AI-enabled tools in the defense sector is especially damaging given the destructive capabilities of the end user as well as the potentially vulnerable proprietary information which can be lost to data leaks.

SCENARIO 1: IMAGE RECOGNITION

Defensive applications of AI technologies have primarily focused on computer vision, a subfield of AI relating to object identification in moving or stationary images. Rapid analysis of large amounts of image data can help individuals in Imagery intelligence classify objects of interest at a faster rate. Project Maven highlights the defense sector's latest attempt to implement computer vision for improved image recognition in combat zones.[1]

Government developers suggest threat actors can manipulate input data which can lead the system to make a false identification. Security researchers have already demonstrated this ability to spoof images, adding strategically placed stickers to a stop sign which led the vehicle's the object detection system to identify the stop sign as a 45mph sign.[2] Attackers can also use strategically placed accessories to fool facial recognition tools used for surveillance purposes.[3] Using compromised object detection systems can have destructive effects if they are implemented in combat zones for defensive purposes as systems may misidentify objects while under duress leading to inappropriate interactions (see Figure 5).

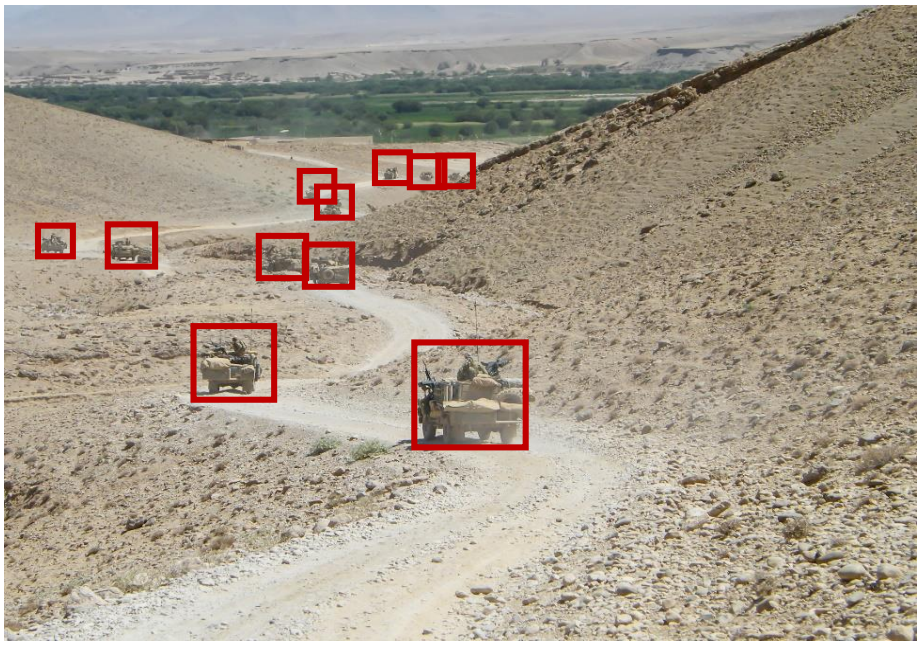


Figure 5. Friendly forces potentially misidentified as enemy combatants.

[1]US Department of Defense, https://www.defense.gov/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/
[3]Eykholt et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, https://arxiv.org/pdf/1707.08945.pdf , 2018
[3]Zhou et al., *Invisible Mask: Practical Attacks on Face Recognition with Infrared*, 2018

SCENARIO 2: NATURAL LANGUAGE PROCESSING

Warfighters in combat zones must deal with a variety of factors when it comes to mission planning and resource allocation. Virtual assistants, which mimic language patterns to sound human, can help warfighters organize taskings, map out capabilities, and make suggestions for action items in demanding environments.[1] However, researchers suggest some AI systems are vulnerable to data leakage and unintended memorization, resulting in models that can reveal sensitive data to users.[2]

Threat actors with access to the source code of a model can use search algorithms to extract data which was used to train the system. In the case of defensive applications, training data could include proprietary information which would normally require security clearance to access. State sponsored actors could extract sensitive political and military information from these assistants to gain a strategic advantage in future malicious activity.

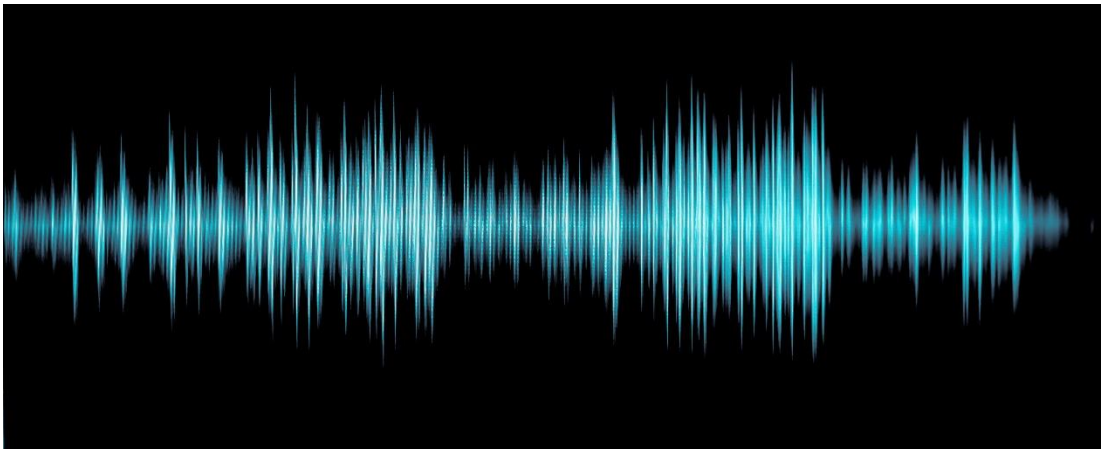


Figure 6. Natural Language Processing

[1]DARPA, https://www.darpa.mil/about-us/timeline/personalized-assistant-that-learns
[2]Carlini et al., *The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets*, https://arxiv.org/pdf/1802.08232.pdf, 2018

SCENARIO 3: PATTERN RECOGNITION

Machine learning allows models trained on data to understand behavioral patterns and help detect anomalies. This capability can help individuals in the counterintelligence sector identify insider threats by analyzing login times, USB inserts, file browsing, and web access activity to detect aberrant behavior and identify potential insider threats.

Malicious actors can leverage data poisoning attacks to manipulate the model by compromising the data used to train the model.[1] This results in a system that is less accurate and will improperly identify insider threats, or neglect to identify genuine insider threats at all. Malicious actors can leverage data poisoning attacks against a counterintelligence sector's AI system to mask the behavior of an insider threat to perpetuate espionage activity.

Figure 7. Recognizing a pattern or anomaly amid baseline noise

[1]Gonzalez et al., *Towards Poisoning of Deep Learning Algorithms with Back-gradient, Optimization*, https://arxiv.org/pdf/1708.08689.pdf, 2017

VULNERABILITIES & RISKS

The three scenarios show how an adversary could potentially use spoofing, data extraction, and poisoning of training data to exploit vulnerabilities and create adverse negative security impacts. Although not comprehensive, the list below provides additional examples of AI vulnerabilities and risks.

Human vulnerabilities. Without education and retraining of the workforce to match the pace of technology changes and the different types of threats, adversaries will encounter fewer obstacles when attempting to exploit AI vulnerabilities. An educated population would also reduce unintended errors as well.

AI is a dual-use technology. AI capabilities possess both civilian and military applications. As warfighters gain more access to AI-enabled tools, malicious actors also find means to leverage AI capabilities for nefarious activity. Recommendation systems that assist warfighters with mission planning and resource allocation can also help attackers choose vulnerable targets for future campaigns.

Data Integrity. AI systems are trained on data to improve the performance of a model. If attackers can figure out the data set used to train a model, they can insert corrupt data that degrades the performance of the final model. If an object detection system used by warfighters is compromised with this type of attack, it can misidentify certain objects (i.e. labeling an enemy tank as a tree), or the system may fail to detect certain objects altogether.

AI applications in irregular warfare for non-state actors. AI is an asymmetric threat, empowering individuals to rival entire states. It is a force multiplier; evolution is fast, completely distributed, and barrier to entry is effectively zero. The AI community's openness and willingness to share ideas leads to the rapid spread of knowledge and resources, which can fall into the hands of threat actors.

STANDARDS TO MITIGATE RISKS

The use of AI can introduce novel, unresolved vulnerabilities that are distinct and unique from traditional information technology systems, as illustrated by the three scenarios. As AI technologies become more widespread, efforts to ensure that they work as intended become more critical.

Our interviews and literature indicated that the application of a standard or combination of standards—such as analytic, research, legal, regulatory, moral, ethical, technical, industry, data, and information security— can help to reduce the risk of adversary exploitation. We took a heuristic approach to tackle a complex problem and selected a sampling of relevant standards in Figure 8, including a description of how it is applicable to AI and how it can reduce AI risk.

WHAT DO WE MEAN BY STANDARDS?

We view the purpose of standards is to facilitate interoperability, identify bias or deviation from a norm or baseline, help measure progress toward a goal, and set uniform requirements. The Institute of Electrical and Electronics Engineers defines standards as "documents that establish specifications and procedures designed to ensure the reliability of the materials, products, methods, and/or services people use every day [to] ensure product functionality and compatibility, facilitate interoperability and support consumer safety and public health." [2]

[1] IEEE, https://beyondstandards.ieee.org/general-news/what-are-standards-why-are-they-important/

STANDARDS TO MITIGATE RISKS (CONTINUED)

Figure 8. Sample Standards Matrix

Type of Standard	How applicable to AI?	Where are the standards applied?	How can it reduce AI risk from an adversary?
Analytic & Research	Standards that evaluate the quality of analysis and scrutability of algorithms	Back end: explainability and transparency	• Identify faulty logic or reasoning, increase the difficulty of deceiving and/or manipulating analysis from AI • Determine how much to trust system inputs and outputs
Legal & Regulatory	Standards based on governance and regulatory oversight into preserving privacy and consent	Front end: usability and personalization Back end: standardized architecture	• Change understanding of liability for mistakes and enhance attribution • Transform notion of jury of peers and evolve crime and punishment
Moral & Ethical	Standards that prevent AI from performing actions that are contrary to a moral or ethical norm	Back end: failsafes	• Reduce likelihood that AI will do the "wrong thing" (i.e. immoral or unethical behavior) if exploited or infiltrated by an adversary
Technical & Industry	Standards to measure the performance of an algorithm on relevant tasks	Front end: performance	• Meet appropriate technical specifications (e.g. low number of false positives) to be robust against adversary denial and deception activities
Data & Information Security	Standards for the protection, sharing, or use of data relevant to a task	Front end: training Back end: data integrity and availability	• Limiting access to and information about how an AI system works to appropriate people could help prevent exploitation by an adversary • Preventing manipulation of training data

A PLACE TO START

The potential demand, whether from government or the general public, that developers produce AI systems that meet agreed-upon standards would help ensure a baseline level of ethics, performance, transparency, etc. for systems that have national security implications. These standards could also help developers understand user requirements and help users make meaningful comparisons between the performances of different AI systems. Academia, industry, and the government already use a number of standards to help measure performance and gauge technological progress. Some examples that are currently used or could be adapted for national security contexts include:

- The MLPerf effort aims to build a common set of benchmarks that enables the machine learning field to measure system performance for both training and inference from mobile devices to cloud services. Their approach is to select a set of machine learning problems—including vision, language, and reinforcement learning—then measure the wall clock time to train a model for each problem.[1] This effort is supported by a number of prominent companies and academic institutions.

- The standard datasets and conditions that are part of open challenges could form the basis for a technical standard adopted on an application-by-application basis. Current examples of such effort include SpaceNet for automatically detecting and extracting features from satellite imagery and NIST's Face Recognition Vendor Test to help measure the accuracy and speed of one-to-many face recognition identification algorithms.[2,3]

[1]MLPerf, https://mlperf.org/
[2]SpaceNet Challenge, http://explore.digitalglobe.com/spacenet
[3]NIST, https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-1n-2018-evaluation

A PLACE TO START (CONTINUED)

- Existing non-AI-related standards could also be updated to reflect new realities brought on by machine learning. The Intelligence Community's Intelligence Community Directive (ICD) 203 analytic tradecraft standards is an example approach. ICD 203 aims to ensure our nation's leaders receive unbiased and accurate intelligence to inform their decisions.

    - ICD 203 governs the production and evaluation of analytic products, as well as, articulates the responsibility of intelligence analysts to strive for excellence, integrity and rigor in their analytic thinking and work practices.[1]

    - The work done as part of DARPA's Explainable AI (XAI) program could be used as a basis for updating ICD 203 Standard 6 – Logical Argumentation.[2] The XAI program seeks to develop new machine-learning systems that will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.

- A number of groups and individuals have called for standards to help ensure that AI systems are aligned to human moral values and ethical principles.[3,4] For example, a 2016 publication by the IEEE, "Ethically Aligned Design," states that AI and autonomous systems have to behave in a way that is beneficial to people beyond reaching functional goals and addressing technical problems.[5] Inclusion of ways to prevent "immoral" or "unethical" behavior could stop AI from doing the wrong thing if exploited or infiltrated by an adversary.

[1]ODNI, https://www.intelligence.gov/mission/our-values/342-objectivity
[2]DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence
[3]Future of Life Institute, https://futureoflife.org/ai-principles/
[4]European Parliament, Artificial Intelligence: Potential Benefits and Ethical Considerations, http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf
[5]IEEE, Ethically Aligned Design, http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf, 2016.

OUTLOOK: A DIALOGUE

Through our research and interviews, we found that identifying standards for AI is a daunting task as AI involves rapidly changing applications and technology that touches almost every sector of society. We are still in the very early stages of understanding and discussing standards for AI. Consequently, we chose the following questions and thoughts on AI and standards to spur discussion to move towards a better understanding of how standards can apply to AI.

Approaches to AI. Altering our mentality to think of AI as a utility, such as *electricity*, will help policymakers better understand how to create standards and regulations. Like AI, electricity can be applied for both nefarious and beneficial purposes. Another approach is to view AI like the *internet,* which has expanded with open published research and cooperation. Finally, examine *other industries*, such as the financial industry for potential best practices on how to implement standards (e.g. capital reserve requirements).

Moral Standards. For AI systems to be used in the service of society, they will need to make recommendations or decisions that align with ethical norms and values. Given that machines are intended to achieve goals with ruthless efficiency, how do we create synergistic or positive relationships between AI and human beings? Moreover, is there an agreed upon threshold of incorporating norms and values without them conflicting with each other?[1]

Transparency. If transparency in algorithmic decision-making became standards for all AI applications, it would defeat many of the national security and homeland security goals given the sensitivity of the information being used. What are the viable pathways to share government information without informing adversaries? What are the criteria to be used to measure quality and explain-ability? How much explain-ability is necessary for a particular function? What impact will an opaque system have on the user?

Performance standards. Are performance standards the appropriate approach to ensure that AI applications work as advertised and are robust against adversary denial and deception activities and other attempts to fool the system? Is the burden to define those performance standards on the implementer or end-user, or on the developer of the system?

[1]For more discussion, see IEEE's *ETHICALLY ALIGNED DESIGN, A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, December 13, 2016

WAY FORWARD: DECISION GUIDE ON AI STANDARDS

We created a decision guide as one starting point for discussion on how to begin to identify, adopt, and implement specific standards for specific needs in the national and homeland security context. While not meant to be exhaustive, we hope these steps will be the building blocks for a public-private sector dialogue on AI and standards.

1) **Determine the national/homeland security context**
- How could your system directly or indirectly impact national or homeland security?
- How does the application fit into the priorities described in documents such as the National Security Strategy of the United States of America, the National Defense Strategy of the United States of America, and the Quadrennial Homeland Security Review?

2) **Identify the type of AI application used**
- What does your AI system do?
- What problem is it meant to solve?

3) **Identify the potential risks and issues associated with the application**
- Through what means could an adversary exploit the system?
- How could an adversary get the system to make a bad decision?
- What critical aspects of the decision-making process are susceptible to adversary exploitation?

4) **Determine the most important metrics that would help indicate that the system is "working in a trustworthy, accurate, appropriate, etc." way**
- How would you be able to tell if the system was working properly?
- How would you measure algorithmic confidence levels?

5) **Choose a type of standard and adjust to the previously identified metrics**
- What standards could address the issues identified in #3?
- What measures would help give you confidence in the results of the algorithm?

6) **Identify key considerations, limitations, and assumptions**
- What aspects of the issue are not addressed by the standard?
- Under what conditions would the standard be effective or ineffective?
- How could an adversary defeat the standards?
- How much would the standards help?
- How much would they hinder development or reduce US competitiveness?

AI TEAM MEMBERS

TERBIUM LABS

Name	Organization
James Sung (Co-Champion)	DHS
Tao Nguyen (Co-Champion)	DIA
Michelle Cantos	FireEye, Inc.
Chandra Pauline Daniel	National Black Leadership Commission on AIDS
Monica Kronauge	DHS/U.S. Customs and Border Protection
Alisa Paige Mason	Guidepost Solutions LLC
JoAnn Ugolini	Hillard Heintze
Munish Walther-Puri	Terbium Labs
Emma Westerman	RAND Corporation

Disclaimer Statement. This document is provided for educational and informational purposes only. The views and opinions expressed in this document do not necessarily state or reflect those of the U.S. Government (USG) or the Public-Private Analytic Exchange Program Partners, and they may not be used for advertising or product endorsement purposes. All judgments and assessments are solely based on unclassified sources and are the product of joint public and USG efforts.

This page intentionally left blank.

This page intentionally left blank.

2018
PUBLIC-PRIVATE
ANALYTIC EXCHANGE PROGRAM