Key Considerations for Responsible Development & Fielding of Artificial Intelligence

July 22, 2020

NATIONAL
SECURITY
COMMISSION
ON ARTIFICIAL
INTELLIGENCE

Key Considerations as a Paradigm for Responsible Development and Fielding of Artificial Intelligence:

National Security Commission on Artificial Intelligence
Line of Effort on Ethics and Responsible AI
Quarter 2 Report: July 22, 2020

Prefatory Note:

This document is a critical excerpt from the National Security Commission on Artificial Intelligence's (NSCAI) Second Quarter Recommendations. The paradigm and recommended practices described here stem from the Commission's line of effort dedicated to Ethics and Responsible AI. The Commission has recommended that heads of departments and agencies critical to national security (at a minimum, the Department of Defense, Intelligence Community, Department of Homeland Security, Federal Bureau of Investigation, Department of Energy, Department of State, and Department of Health and Human Services) should implement the Key Considerations as a paradigm for the responsible development and fielding of AI systems. This includes developing processes and programs aimed at adopting the paradigm's recommended practices, monitoring their implementation, and continually refining them as best practices evolve.

This approach would set the foundation for an intentional, government-wide, coordinated effort to incorporate recommended practices into current processes for AI development and fielding. However, our overarching aim is to allow agencies to continue to have the flexibility to craft policies and processes according to their specific needs. The Commission is mindful of the required flexibility that an agency needs when conducting the risk assessment and management of an AI system, as these tasks will largely depend on the context of the AI system.

This recommendation along with a set of recommended considerations and practices was made in July 2020 as part of the Commission's broader recommendations to the Executive and Legislative Branches, found here.

We note that an abridged version of this document is available.

Key Considerations for Responsible Development & Fielding of Artificial Intelligence

Introduction

In the Commission's Interim Report, we stated that "defense and national security agencies must develop and deploy AI in a responsible, trusted, and ethical manner to sustain public support, maximize operational effectiveness, maintain the integrity of the profession of arms, and strengthen international alliances."[1]

As the Commission makes recommendations to advance ethical and responsible AI for national security, we are aware that this topic presents unique challenges. Concerns about the responsible development and fielding of AI technologies span a range of issues. Many debates are ongoing as the technology and its applications rapidly evolve, and the need for norms and best practices becomes more apparent.

The Commission acknowledges the efforts undertaken to date to establish ethics guidelines for AI by entities in government, in the private sector, and around the world.[2] The Department of Defense took the critical step of adopting a set of high-level principles to guide its development and use of AI.[3] While some agencies critical to national security have adopted, or are in the process of adopting, AI principles,[4] others agencies have not provided such guidance. In cases where principles are offered, it can be difficult to translate the high-level concepts into concrete actions. There is often a gap between articulating high-level goals around responsible AI and operationalizing them.

In addition, agencies would benefit from the establishment of greater consistency in policies to further the responsible development and fielding of AI technologies across government. A unified approach would not only be more efficient, but it could also stimulate innovation and efficiencies through the sharing of models, data, and other information. Below the Commission is identifying a set of challenges and making recommendations on directions with responsibly developing and fielding AI systems,

[1] *Interim Report,* NSCAI at 16 (Nov. 2019), https://www.nscai.gov/reports [hereinafter Interim Report].

[2] Examples of efforts to establish ethics guidelines are found within the U.S. government, industry, and internationally. See e.g., *Draft Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*, Office of Management and Budget (Jan. 1, 2019), https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf; Jessica Fjeld & Adam Nagy, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center (Jan. 15, 2020), https://cyber.harvard.edu/publication/2020/principled-ai; *OECD Principles on AI*, OECD (last accessed June 17, 2020), https://www.oecd.org/going-digital/ai/principles/; *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, European Union at 26-31 (Apr. 8, 2019), https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

[3] C. Todd Lopez, *DOD Adopts 5 Principles of Artificial Intelligence Ethics*, Department of Defense (Feb. 5, 2020), https://www.defense.gov/Explore/News/Article/Article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/ [hereinafter, Lopez, DOD Adopts 5 Principles].

[4] See Ben Huebner, *Presentation: AI Principles*, Intelligence and National Security Alliance 2020 Spring Symposium, Building an AI Powered IC, (Mar. 4, 2020), https://www.insaonline.org/2020-spring-symposium-building-an-ai-powered-ic-event-recap/.

and for pinpointing the concrete actions that should be adopted across the government to help overcome these challenges.

This Commission has assessed a set of recommended practices in five categorical areas that are ripe for adoption. Collectively, they form a paradigm for aligning AI system development and AI system behavior to goals and values. The first section provides guidance specific to implementing systems that abide by American values and the rule of law. The section covers aligning the run-time behavior of systems to the related, more technical encodings of objectives, utilities, and trade-offs. The four following sections (on *Engineering Practices, System Performance, Human-AI Interaction*, and *Accountability & Governance*) serve in support of core American values and outline practices needed to develop and field systems that are trustworthy, understandable, reliable, and robust. Recommended practices span multiple phases of the *AI lifecycle*, from conception and early design, through development and testing, and maintenance and technical refresh. The Commission uses "development" to refer to 'designing, building, and testing during development and prior to deployment' and "fielding" to refer to 'deployment, monitoring, and sustainment.'

Though best practices will evolve (for instance, through future R&D), these recommended practices establish a baseline for the responsible development and fielding of AI technologies. They provide a floor, rather than a ceiling, for the responsible development and fielding of AI technologies. The Commission recommends that heads of departments and agencies implement the Key Considerations as a paradigm for the responsible development and fielding of AI systems. This includes developing processes and programs aimed at adopting the paradigm's recommended practices, monitoring their implementation, and continually refining them as best practices evolve. These practices imply derived requirements for AI systems, requirements that in turn become an integral part of an agency's risk management process when deciding whether and how to develop and use AI for the context at hand. These recommended practices should apply both to systems that are developed by departments and agencies, as well as those that are acquired. Systems acquired (whether commercial, off-the-shelf systems or those acquired through contractors) should be subjected to the same rigorous standards and practices—whether in the acquisitions or acceptance processes. As such, the government organization overseeing the bidding process should require assertions of goals aligned with recommended practices for the Key Considerations in the process.

In each of the five categorical areas that follow, we first provide a conceptual overview of the scope and importance of the topic. We then illustrate an example of a current challenge relevant to national security departments that underscores the need to adopt recommended practices in this area. Then, we provide a list of recommended practices that agencies should adopt, acknowledging research, industry tools, and exemplary models within government that could support agencies in the adoption of recommended practices. Finally, in areas where recommended practices do not exist or they are especially challenging to implement, we note the need for future work as a priority; this includes, for example, R&D and standards development. We also identify potential areas in which collaboration with allies and partners would be beneficial for interoperability and trust, and note that the Key Considerations can inform potential future efforts to discuss military uses of AI with strategic competitors.

I. Aligning Systems and Uses with American Values and the Rule of Law

(1) Overview:

Our values guide our decisions and our assessment of their outcomes. Our values shape our policies, our sensitivities, and how we balance trade-offs among competing interests. Our values, and our commitment to upholding them, are reflected in the U.S. Constitution, and our laws, regulations, programs, and processes.

One of the seven principles we set forth in our Interim Report (November 2019) is the following:

> The American way of AI must reflect American values—including having the rule of law at its core. For federal law enforcement agencies conducting national security investigations in the United States, that means using AI in ways that are consistent with constitutional principles of due process, individual privacy, equal protection, and non-discrimination. For American diplomacy, that means standing firm against uses of AI by authoritarian governments to repress individual freedom or violate the human rights of their citizens. And for the U.S. military, that means finding ways for AI to enhance its ability to uphold the laws of war and ensuring that current frameworks adequately cover AI.[5]

Values established in the U.S. Constitution, and further operationalized in legislation, include freedoms of speech and assembly, the rights to due process, inclusion, fairness, non-discrimination (including equal protection), and privacy (including protection from unwarranted government interference in one's private affairs).[6] Beyond the values codified in the U.S. Constitution and the U.S. Code, our values also are expressed via international treaties that the United States has ratified that affirm our commitments to human rights and human dignity, including the International Convention of Civil and Political Rights.[7] Within America's national

[5] *Interim Report* at 17.

[6] See e.g., U.S. Const. amendments I, IV, V, and XIV; Americans with Disability Act of 1990, 42 U.S.C. § 12101 et seq.; Title VII of the Consumer Credit Protection Act, 15 U.S.C. §§ 1691-1691f; Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq..

[7] International Covenant on Civil and Political Rights, UN General Assembly, United Nations, Treaty Series, vol. 999 at 171 (Dec. 16, 1966), https://www.refworld.org/docid/3ae6b3aa0.html. As noted in the Commission's Interim Report, America and its like-minded partners share a commitment to democracy, human dignity and human rights. See Interim Report at 48. Many, but not all nations,

security departments, our commitment to protecting and upholding privacy and civil liberties is further embedded in the policies and programs of the Intelligence Community,[8] the Department of Homeland Security,[9] the Department of Defense (DoD),[10] and oversight entities.[11] This is not an exhaustive set of values that U.S. citizens would identify as core principles of the United States. However, the paradigm of considerations and recommended practices for AI that we introduce resonate with these highlighted values as they have been acknowledged and elevated as critical by the U.S. government and national security departments and agencies. Further, many of these values are common to America's like-minded partners who share a commitment to democracy, human dignity, and human rights.

In the military context, core values such as distinction and proportionality are embodied in the nation's commitment to, and the DoD's policies to uphold, the Uniform Code of Military Justice and the Law of Armed Conflict (LOAC).[12] Other values are reflected in treaties, rules, and policies such as the Convention Against

share commitments to these values. Even when values are shared, however, they can be culturally relative, for instance, across nations, owing to interpretative nuances.

[8] See e.g., Daniel Coats, *Intelligence Community Directive 107*, ODNI (Feb. 28, 2018), https://fas.org/irp/dni/icd/icd-107.pdf (on protecting civil liberties and privacy); *IC Framework for Protecting Civil Liberties and Privacy and Enhancing Transparency Section 702*, Intel.gov (Jan. 2020), https://www.intelligence.gov/index.php/ic-on-the-record/guide-to-posted-documents#SECTION_702-OVERVIEW (on privacy and civil liberties implication assessments and oversight); *Principles of Professional Ethics for the Intelligence Community*, ODNI, (https://www.dni.gov/index.php/who-we-are/organizations/clpt/clpt-related-menus/clpt-related-links/ic-principles-of-professional-ethics (last visited June 17, 2020) (on diversity and inclusion).

[9] See e.g., *Privacy Office*, Department of Homeland Security (June 3, 2020), https://www.dhs.gov/privacy-office#; *CRCL Compliance Branch*, Department of Homeland Security (May 15, 2020), https://www.dhs.gov/compliance-branch.

[10] See Samuel Jenkins & Alexander Joel, *Balancing Privacy and Security: The Role of Privacy and Civil Liberties in the Information Sharing Environment*, IAPP Conference 2010 (2010), https://dpcld.defense.gov/Portals/49/Documents/Civil/IAPP.pdf.

[11] See *Projects*, U.S. Privacy and Civil Liberties Oversight Board, (last visited June 17, 2020), https://www.pclob.gov/Projects.

[12] See *Department of Defense Law of War Manual*, DoD Office of General Counsel (Dec. 2016), https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190 [hereinafter DoD Law of War Manual]. See also *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document*, Defense Innovation Board (Oct. 31, 2019), https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF ("More than 10,000 military and civilian lawyers within DoD advise on legal compliance with regard to the entire range of DoD activities, including the Law of War. Military lawyers train DoD personnel on Law of War requirements, for example, by providing additional Law of War instruction prior to a deployment of forces abroad. Lawyers for a Component DoD organization advise on the issuance of plans, policies, regulations, and procedures to ensure consistency with Law of War requirements. Lawyers review the acquisition or procurement of weapons. Lawyers help administer programs to report alleged violations of the Law of War through the chain of command and also advise on investigations into alleged incidents and on accountability actions, such as commanders' decisions to take action under the Uniform Code of Military Justice. Lawyers also advise commanders on Law of War issues during military operations.").

Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment;[13] the DoD's Rules of Engagement;[14] and the DoD's Directive 3000.09.[15]

U.S. values demand that the development and use of AI respect these foundational values, and that they enable human empowerment as well as accountability. They require that the operation of AI systems and components be compliant with our laws and international legal commitments, and with departmental policies. In short, core American values must inform the way we develop and field AI systems, and the way our AI systems behave in the world.

To date, AI Principles adopted and endorsed by the Executive Branch, including by national security department and agencies, have focused on aligning AI with many of the values discussed in this section, including fairness and non-discrimination,[16] privacy and civil liberties,[17] and accountability.[18] Taking the DoD Principles as one example, fairness is evoked by the "Equitable" principle that the department will "take deliberate steps to minimize unintended bias in AI capabilities."[19] Accountability is evoked by the "Responsible" principle that "DoD personnel will exercise appropriate levels of judgment and care while remaining responsible for the development, deployment and use of AI capabilities."[20] The work on establishing principles reiterates the importance of developing and deploying AI systems in accordance with these values. They form the foundation that the Commission's recommendations build upon.

(2) Examples of Current Challenges

Machine learning techniques can assist DoD agencies with conducting large-scale data analyses to support and enhance decision-making about personnel. As an example, the JAIC Warfighter Health Mission Initiative Integrated Disability Evaluation System model seeks to leverage data analyses to identify service members

[13] Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, United Nations General Assembly (Dec. 10, 1984), https://www.ohchr.org/en/professionalinterest/pages/cat.aspx.

[14] See DoD Law of War Manual at 26 (("Rules of Engagement reflect legal, policy, and operational considerations, and are consistent with the international law obligations of the United States, including the law of war.").

[15] See *Department of Defense Directive 3000.09 on Autonomy in Weapons Systems*, Department of Defense (Nov. 21 2012), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf ("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.").

[16] See e.g., Lopez, DOD Adopts 5 Principles; *Draft Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*, Office of Management and Budget (Jan. 1, 2019), https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

[17] See Ben Huebner, *Presentation: AI Principles*, Intelligence and National Security Alliance 2020 Spring Symposium, Building an AI Powered IC (Mar. 4, 2020), https://www.insaonline.org/2020-spring-symposium-building-an-ai-powered-ic-event-recap/.

[18] Id.

[19] See Lopez, DOD Adopts 5 Principles.

[20] Id.

on the verge of ineligibility due to concerns with their readiness[21]. Other potential analyses can support personnel evaluations, including analyzing various factors that lead to success or failure in promotion. Caution and proven practices are needed however to avoid pitfalls in fairness and inclusiveness, several of which have been highlighted in high-profile challenges in such areas as criminal justice,[22] recruiting and hiring,[23] and face recognition.[24] Attention should be paid to challenges with decision support systems to avoid harmful disparate impact.[25] Likewise, factors chosen to weigh in performance evaluations and promotions must be carefully considered to avoid inadvertently reinforcing existing biases through ML-assisted decisions.

(3) Recommendations for Adoption

Recommended Practices to Implement American Values

A. Developing uses and building systems that behave in accordance with American values and the rule of law.

1. **Employ technologies and operational policies that align with privacy preservation, fairness, inclusion, human rights, and law of armed conflict.** Technologies and policies throughout the AI lifecycle should support achieving the goals that AI uses and systems are consistent with these values—and should mitigate the risk that AI system uses/outcomes

[21] See *JAIC Mission Initiative in the Spotlight: Warfighter Health*, JAIC (Apr. 15, 2020), https://www.ai.mil/blog_04_15_20-jaic_mi_warfighter_health.html.

[22] *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System,* Partnership on AI, (last accessed July 14, 2020), https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/.

[23] Andi Peng et al., *What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring*, Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (Oct. 2019), https://arxiv.org/pdf/1909.03567.pdf; Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [hereinafter Dastin. Amazon Scraps Secret AI Recruiting Tool].

[24] Patrick Grother, et. al., *Face Recognition Vendor Test (FRVT) Part Three: Demographic Effects*, National Institute of Standards and Technology (Dec. 2019), https://doi.org/10.6028/NIST.IR.8280 [hereinafter Grother, Face Recognition Vendor Test (FRVT) Part Three: Demographic Effects].

[25] PNDC provides predictive analytics to improve military readiness; enable earlier identification of service members with potential unfitting, disabling, or career-ending conditions; and offer opportunities for early medical intervention or referral into disability processing. To do so, PNDC provides recommendations at multiple points in the journey of the non-deployable service member through the Military Health System to make "better decisions" that improve medical outcomes and delivery of health services. This is very similar to the OPTUM decision support system that recommended which patients should get additional intervention to reduce costs. Analysis showed millions of US patients were processed by the system, with substantial disparate impact on black patients compared to white patients. Shaping development from the start to reflect bias issues (which can be subtle) would have produced a more equitable system and avoided scrutiny and suspension of system use when findings were disclosed. See Heidi Ledford, *Millions of Black People Affected by Racial Bias in Health Care Algorithms*, Nature (Oct. 26, 2019), https://www.nature.com/articles/d41586-019-03228-6.

will violate these values. While not an exhaustive list, we offer the following examples based upon core values discussed above:

- For ensuring *privacy*, employ privacy protections, privacy-sensitive analyses, eyes-off ML, ML with encrypted data and models, and multi-party computation methods.
- For *fairness and to mitigate unwanted bias*, use tools to probe for unwanted bias in data, inferences, and recommendations. [26]
- For *inclusion*, ensure usability of systems, accessible design, appropriate ease of use, learnability, and training availability.
- For commitment to *human rights,* place limitations and constraints on applications that would put commitment to human rights at risk, for example, limits on storing observational data beyond its specific use or using data for purposes other than its primary, intended focus.
- For compliance with the *Law of Armed Conflict*, tools for interpretability and to provide cues to the human operator should enable context-specific judgments to ensure, for instance, distinction between active combatants, those who have surrendered, and civilians.[27]

B. **Representing Objectives and Trade-offs**

Above, we described the goals of developing systems that align with key values through employing technologies and operational policies. Another important practice for aligning AI systems with values is to consider values as (1) embodied in choices about engineering trade-offs and (2) explicitly represented in the goals and utility functions of an AI system.[28]

On (1), multiple trade-offs may be encountered with the engineering of an AI system. With AI, trade-offs need to be made based on what is most valued (and the benefits and risks to those values)[29] including for high-stakes, high-risk pattern

[26] Data should be appropriately biased (in a statistical sense) for what it's needed to do in order to have accurate predictions. However, beyond this, diverse concerns with unwanted bias exist, including factors that could make a system's outcomes morally or legally unfair. See Ninaresh Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning*, USC Information Sciences Institute (Sept. 17, 2019) https://arxiv.org/pdf/1908.09635.pdf. For an illustration of ways fairness can be assessed across the AI lifecycle, see Sara Robinson, *Building Machine Learning Models for Everyone: Understanding Fairness in Machine Learning*, Google (Sept. 25, 2019) https://cloud.google.com/blog/products/ai-machine-learning/building-ml-models-for-everyone-understanding-fairness-in-machine-learning.

[27] For more examples on the law of armed conflict, see *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach*, International Committee of the Red Cross (June 6, 2019), https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach.

[28] Mohsen Bayati, et al., *Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study*, PLOS One Medicine (Oct. 2014), https://doi.org/10.1371/journal.pone.0109264; Eric Horvitz & Adam Seiver, *Time-Critical Action: Representations and Application*, Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (Aug. 1997), https://arxiv.org/pdf/1302.1548.pdf.

[29] Jessica Cussins Newman, *Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles*, Berkeley Center for Long-Term Cybersecurity (May 5, 2020),

recognition, recommendation, and decision making under uncertainty. Trade-off decisions for AI systems must be made about internal representations, policies of usage and controls, run-time execution monitoring, and thresholds. These include a number of well-known, inescapable engineering trade-offs when it comes to building and using machine-learning to develop models for prediction, classification, and perception. For example, systems that perform recognition or prediction tasks can be set to work at different operating thresholds or settings (along a well-characterized curve) where different settings change the trade-off between precision and recall or the rates of true positives and false positives. By changing the settings, the ratio of true positives to false positives is changed. Often, one can raise the rate of true positives but will also raise the false negatives.[30] In high-stakes applications, different kinds of inaccuracies (e.g., missing a recognition and falsely recognizing) are associated with different outcomes and costs. Thus, the setting of thresholds and understanding the influences of different settings on the behavior of a system *entail making value judgments*. As with all engineering trade-offs, making choices about trade-offs explicitly and deliberately provides more transparency, accountability, and confidence in the process than making decisions implicitly and ad hoc as they arise.

On (2), systems may be guided by optimization processes that seek to maximize an objective function or *utility model*. Such objectives can represent a set of independent goals, as in *multi-objective optimization*. A multi-attribute utility function may be employed to guide a system's actions based on an objective that is constructed by weighing several individual factors, where explicit weights are assigned to capture the asserted importance of each of the different factors. Different weightings on factors can be viewed as embedding different values into a system. Here too trade-offs are made when using multi-attribute utility or objective functions within applications.[31] Even when tuning a model for fairness, when multiple metrics of fairness are relevant, optimizing for one metric can cause a trade-off in performance across the second metric.[32] As a result, it is important to acknowledge inherent trade-offs and the need for setting or encoding "preferences" - which requires *someone* or *some organization* to make a call

https://cltc.berkeley.edu/ai-decision-points/ [hereinafter Newman, Decision Points in AI Governance].

[30] For more on the trade-offs between false positive and false negative rates, and the implications of chosen thresholds, see Grother, Face Recognition Vendor Test (FRVT) Part Three: Demographic Effects.

[31] Optimal decisions may require making a decision when trade-offs exist between two or more conflicting objectives. For example, a predictive maintenance system for aircraft will have objectives that are in tension including: minimizing false positives, minimizing false negatives, minimizing the need for instrumentation on the aircraft, maximizing the specificity of the recommended maintenance action, and adapting to new operational profiles the aircraft perform in over time.

[32] It is sometimes impossible to simultaneously satisfy different fairness criteria. See Yungfeng Zhang, et al., *Joint Optimization of AI Fairness and Utility: A Human-Centered Approach, Association for Computing Machinery*, AIES '20 (Feb. 7-8, 2020), https://dl.acm.org/doi/10.1145/3375627.3375862.

about the right trade.[33]

Recommended Practices for Representing Objectives and Trade-offs

1. **Consider and document value considerations in AI systems and components based on specifying how trade-offs with accuracy are handled;** this includes operating thresholds that yield different true positive and false positive rates or different precision and recall.
2. **Consider and document value considerations in AI systems that rely on representations of objective or utility functions,** including the handling of multi-attribute or multi-objective models.
3. **Conduct documentation, reviews, and set limits on disallowed outcomes.** It is important to:
   - Be transparent and keep documentation on assertions about the trade-offs made, optimization justifications, and acceptable thresholds for false positives and false negatives.
   - During system development and testing, consider the potential need for context-specific changes in goals or objectives that would require a revision of parameters on settings or weightings on factors.
   - Establish explicit controls in specific use cases and have the capability to change or set controls, potentially by context or by policy, per organization.
   - Review documentation and run-time execution trade-offs, potentially on a recurrent basis, by appropriate experts/authorities.
   - Acknowledge that performance characteristics are statistics over multiple cases, and that different settings and workloads have different performance.
   - Set logical limits on disallowed outcomes, where needed, to put additional constraints on allowed performance.

(4) Recommendations for Future Action

Future R&D is needed to advance capabilities for preserving and ensuring that developed or acquired AI systems will act in accordance with American values and the rule of law. For instance, the Commission notes the need for R&D to assure that the personal privacy of individuals is protected in the acquisition and use of data for AI system development.[34] This includes advancing ethical practices with the use of personal data, including disclosure and consent about data collection and use models (including uses of data to build base models that are later retrained and fine-tuned for specific tasks). R&D should also advance development of anonymity techniques and privacy-preserving technologies including homomorphic encryption and differential

[33] See *Analyses of Alternatives, Systems Engineering Guide*, MITRE (May 2014), https://www.mitre.org/publications/systems-engineering-guide/acquisition-systems-engineering/acquisition-program-planning/performing-analyses-of-alternatives.

[34] The Commission is doing a fulsome assessment of where investment needs to be made; this document notes important R&D areas through the lens of ethics and responsible AI.

privacy techniques and identify optimal approaches for specific use cases. Research should focus upon advancing multi-party compute capabilities (to allow collaboration on the pooling of data from multiple organizations without sharing datasets), and developing a better understanding of the compatibility of the promising privacy preserving approaches with regulatory approaches such as the European Union's General Data Protection Regulation (GDPR), as both areas are important for allied cooperation.

II. Engineering Practices

(1) Overview

The government, and its partners (including vendors), should adopt recommended practices for creating and maintaining trustworthy and robust AI systems that are *auditable* (able to be interrogated and yield information at each stage of the AI lifecycle to determine compliance with policy, standards, or regulations[35]); *traceable* (to understand the technology, development processes, and operational methods applicable to AI capabilities, e.g., with transparent and auditable methodologies, data sources, and design procedure and documentation[36]); *interpretable* (to understand the value and accuracy of system output[37]), *and reliable* (to perform in the intended manner within the intended domain of use[38]).

There are no broadly directed best practices or standards (e.g., endorsed by the Secretary of Defense or Director of National Intelligence) in place to define how organizations should build AI systems that are consistent with designated AI principles. But efforts in commercial, scientific, research, and policy communities are generating candidate approaches, minimal standards, and engineering proven practices to ensure the responsible design, development, and deployment of AI systems.[39]

While AI refers to a constellation of technologies, including logic-based systems, the rise in capabilities in AI systems over the last decade is largely attributable to

[35] See Inioluwa Deborah Raji, et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing,* ACM FAT (Jan. 3, 2020), https://arxiv.org/abs/2001.00973 [hereinafter Raji, Closing the AI Accountability Gap].

[36] Lopez, DOD Adopts 5 Principles.

[37] *Model Interpretability in Azure Machine Learning (preview)*, Microsoft (July 2020), https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability.

[38] Lopez, DOD Adopts 5 Principles.

[39] See Newman, Decision Points in AI Governance; Raji, Closing the AI Accountability Gap; Miles Brundage, et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (Apr. 20, 2020), https://arxiv.org/abs/2004.07213 [hereinafter Brundage, Toward Trustworthy AI Development]; Saleema Amershi, et. al., *Software Engineering for Machine Learning: A Case Study*, Microsoft (Mar. 2019), https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf [hereinafter Amershi, Software Engineering for Machine Learning].

capabilities provided by data-centric machine learning (ML) methods. New security and robustness challenges are linked to different phases of ML system construction and operations.[40] Several properties of the methods and models used in ML are associated with weaknesses that make the systems brittle and exploitable in specific ways—and vulnerable to failure modalities not seen in traditional software systems. Such failures can rise inadvertently or as the intended results of malicious attacks and manipulation. Attributes of machine learning training procedures and run-times linked to intentional and unintentional failures include: (1) the critical reliance on data for training, (2) the common use of such algorithmic procedures as differentiation and gradient descent to construct and optimize the performance of models, (3) the ability to probe models with multiple tasks or queries, and (4) the possibility of gaining access to information about models and their parameters.

Given the increasing consequences of failure in AI systems as they are integrated into critical uses, the various failure modes of AI systems have received significant attention. The exploration of AI failure modes has been divided into adversarial attacks[41] or unintended faults introduced throughout the lifecycle.[42] The pursuit of security and robustness of AI systems requires awareness, attention, and proven practices around intentional and unintentional failure modes.[43]

Intentional failures are the result of malicious actors explicitly attacking some aspect of (AI) system training or run-time behavior. Researchers and practitioners in the evolving area of Adversarial Machine Learning *(*AML) have created taxonomies of malicious attacks on machine learning training procedures and run-times. Attacks span ML training and testing and each has associated defenses.[44] Categories of intentional failures introduced by adversaries include training *data poisoning* attacks, *model inversion*, and ML *supply chain attacks*.[45] National security uses of AI are likely targets of sustained adversarial efforts; awareness of sets of potential vulnerabilities and proven practices for detecting attacks and protecting systems is critical. AI developed for this community must remain current with a rapidly developing

[40] Elham Tabassi, et al., *A Taxonomy and Terminology of 4 Adversarial Machine Learning (Draft NISTIR 8269)*, National Institute of Standards and Technology (Oct. 2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf [hereinafter Tabassi, A Taxonomy and Terminology of 4 Adversarial Machine Learning (Draft NISTIR 8269)].

[41] See Guofu Li, et al., *Security Matters: A Survey on Adversarial Machine Learning*, (Oct. 2018), https://arxiv.org/abs/1810.07339; Tabassi, A Taxonomy and Terminology of 4 Adversarial Machine Learning (Draft NISTIR 8269).

[42] See José Faria, *Non-Determinism and Failure Modes in Machine Learning*. 2017 IEEE 28th International Symposium on Software Reliability Engineering Workshops (Oct. 2017), https://ieeexplore.ieee.org/document/8109300; Dario Amodei et al., *Concrete Problems in AI Safety*, (Jun. 2016), https://arxiv.org/abs/1606.06565.

[43] Ram Shankar Siva Kumar et al., *Failure Modes in Machine Learning,* (Nov. 2019), https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning [hereinafter, Kumar, Failure Modes in Machine Learning].

[44] See Tabassi, A Taxonomy and Terminology of 4 Adversarial Machine Learning (Draft NISTIR 8269).

[45] For 11 categories of attack, and associated overviews, see the "Intentionally-Motivated Failures Summary" in Kumar, Failure Modes in Machine Learning.

understanding of the nature of vulnerabilities to attacks as these attacks grow in sophistication. Advances in new attack methods and vectors must be followed with care and recommended practices implemented around technical and process methods for mitigating vulnerabilities and detecting, alerting, and responding to attacks.

Unintentional failures can be introduced at multiple points in the AI development and deployment lifecycle. In addition to faults that can be inadvertently introduced into any software development effort (e.g., requirements ambiguity, coding errors, inadequate TEVV, flaws in tools used to develop and evaluate the system), distinct additional failure modes can be introduced for machine learning systems. Examples of unintentional AI failures (with particular relevance to deep learning and reinforcement learning) include *reward hacking, side-effects, distributional shifts*, and *natural adversarial examples*.[46] Another area of failure includes the inadequate specification of values per objectives represented in system utility functions (as described in Section 1 above on *Representing Objectives and Trade-offs*), leading to unexpected and costly behaviors and outcomes, akin to outcomes in the fable of the Sorcerer's Apprentice[47]. Additional classes of unintentional failures can arise as unexpected and potentially costly behaviors generated via the interactions of multiple distinct AI systems that are each developed and tested in isolation. The explicit or inadvertent composition of sets of AI systems within one's own services, forces, agencies, and between US systems and those of allies, adversaries, and potential adversaries, can lead to complex multi-agent situations with unexpected and poorly-characterized behaviors.[48]

(2) Examples of Current Challenges

To make high-stakes decisions, and often in safety-critical contexts, DoD and the IC must be able to depend on the integrity and security of the data that is used to train some kinds of ML systems. The challenges of doing so have been echoed by the leadership of the DoD and the Intelligence Community,[49] including concerns with

[46] Id.

[47] Thomas Dietterich & Eric Horvitz, *Rise of Concerns about AI: Reflections and Directions*, Communications of the ACM, Vol. 58 No. 10 at 38-40 (Oct. 2015), http://erichorvitz.com/CACM_Oct_2015-VP.pdf .

[48] Unexpected performance represents emergent runtime output, behavior, or effects at the system level, e.g., through unanticipated feature interaction, … that was also not previously observed during model validation." See Colin Smith et al., *Hazard Contribution Modes of Machine Learning Components*, AAAI-20 Workshop on Artificial Intelligence Safety (SafeAI 2020) (Feb. 7, 2020), https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20200001851.pdf.

[49] See Don Rassler, *A View from the CT Foxhole Lieutenant General John N.T. "Jack" Shanahan, Director, Joint Artificial Intelligence Center, Department of Defense*, Combating Terrorism Center at West Point (Dec. 2019), https://ctc.usma.edu/view-ct-foxhole-lieutenant-general-john-n-t-jack-shanahan-director-joint-artificial-intelligence-center-department-defense/ ("I am very well aware of the power of information, for good and for bad. The profusion of relatively low-cost, leading-edge information-related capabilities and advancement of AI-enabled technologies such as generative adversarial networks or GANs, has made it possible for almost anyone—from a state actor to a lone wolf terrorist—to use information as a precision weapon. What was viewed largely as an annoyance a few years ago has now

detecting adversarial attacks such as data poisoning, sensor spoofing, and "enchanting attacks" (when the adversary lures a reinforcement learning agent to a designated target state that benefits the adversary).[50]

(3) Recommendations for Adoption

Engineering Recommended Practices

Critical engineering practices needed to operationalize AI principles (such as 'traceable' and 'reliable'[51]) are described in the non-exhaustive list below. These practices span design, development, and deployment of AI systems.

1. **Concept of operations development and design and requirements definition and analysis**. Conduct systems analysis of operations and identify mission success metrics. Identify potential functions that can be performed by an AI technology. Incorporate early analyses of use cases and scenario development, assess general feasibility, and make a critical assessment of the reproducibility and demonstrated maturity of specific candidate AI technologies. This includes broad stakeholder engagement and hazard analysis, including domain experts and individuals with expertise and/or training in the responsible development and fielding of AI technologies. This includes for example asking key questions about potential disparate impact early in the development process and documenting deliberations, actions, and approaches used to ensure fairness and lack of unwanted bias in the machine learning application.[52] The feasibility of

become a serious threat to national security. Even more alarming, it's almost impossible to predict the exponential growth of these information-as-a-weapon capabilities over the next few years."); see also Dean Souleles, *2020 Spring Symposium: Building an AI Powered IC*, Intelligence and National Security Alliance (Mar. 9, 2020), https://www.insaonline.org/2020-spring-symposium-building-an-ai-powered-ic-event-recap/ ("We need to be thinking of authenticity of information and provenance of information.....How do you know that the news you are reading is authentic news? How do you know that its source has provenance? How can you trust the information of the world? And in this era of deep fakes and generative artificial neural networks scans that can produce images and texts and videos and audio that are increasingly indistinguishable from authentic, where then is the role of the intelligence officer? If you can no longer meaningfully distinguish truth from falsehood, how do you write an intelligence report? How do you tell national leadership with confidence you believe something to be true or not to be true. That is a big challenge. . . . We need systems that are reliable and understandable. We need to be investing in the gaps.").

[50] Naveed Akhtar & Ajmal Mian, *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey* (Feb. 2018), https://arxiv.org/abs/1801.00553.

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

[52] There is no single definition of fairness. System developers and organizations fielding applications must work with stakeholders to define fairness, and provide transparency via disclosure of assumed definitions of fairness. Definitions or assumptions about fairness and metrics for identifying fair inferences and allocations should be explicitly documented. This should be accompanied by a discussion of alternate definitions and rationales for the current choice. These elements should be

meeting these requirements may trigger a review of whether and where it is appropriate to use AI in the system being proposed. Opportunities exist to use experimentation, modeling/simulation, and rapid prototyping of AI systems to validate operational requirements and assess feasibility.[53]

- **Risk assessment**. In conducting stakeholder engagement and hazard analysis, it is important to assess risks and trade-offs with a diverse interdisciplinary group. This includes a discussion of a system's potential societal impact. Prior to developing or acquiring a system, or conducting AI R&D in a novel area, risk assessment questions should be asked relevant to the national security context in critical areas, including questions about privacy and civil liberties, the law of armed conflict, human rights,[54] system security, and the risks of a new technology being leaked, stolen, or weaponized. [55]

2. **Documentation of the AI lifecycle:** Whether building and fielding an AI system or "infusing AI" into a preexisting system, require documentation[56] on:

documented internally as machine-learning components and larger systems are developed. This is especially important as establishing alignment on the metrics to use for assessing fairness encounters an added challenge when different cultural and policy norms are involved when collaborating on development and use with allies.

[53] Design reviews take place at multiple stages in the Defense Acquisition process. Recent reforms to the Defense Acquisition System efforts, include the release of a new DoD 5000.02, which issues the "Adaptive Acquisition Framework" and an interim policy for a software acquisition pathway; this reflects efforts to further adapt the system to support agile and iterative approaches to software-intensive system development. See *Software Acquisition*, Defense Acquisition University (last visited June 18, 2020), https://aaf.dau.edu/aaf/software/; *DoD Instruction 5000.02: Operation Of The Adaptive Acquisition Framework*, Office of the Under Secretary of Defense for Acquisition and Sustainment (Jan. 23, 2020) https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf?ver=2020-01-23-144114-093.

[54] For more on the importance of human rights impact assessments of AI systems, see *Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression,* UN Human Rights Office of the High Commissioner (2018), https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx. For an example of a human rights risk assessment for AI in categories such as nondiscrimination and equality, political participation, privacy, and freedom of expression, see Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data Society (Oct. 2018),. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

[55] For exemplary risk assessment questions that IARPA has used, see Richard Danzig, *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*, Center for a New American Security at 22 (June 28, 2018), https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101.

[56] Documentation recommendations build off of a legacy of robust documentation requirements. See *Department of Defense Standard Practice: Documentation of Verification, Validation, and Accreditation (VV&A) For Models and Simulations*, Department of Defense (Jan. 28, 2008), https://acqnotes.com/Attachments/MIL-STD-3022%20Documentation%20of%20VV&A%20for%20Modeling%20&%20Simulation%2028%20Jan%2008.pdf.

- If ML is used, the data used for training and testing, including clear and consistent annotation of data, the origin of the data (e.g., why, how, and from whom), provenance, intended uses, and any caveats with re-uses;[57]
- The algorithm(s) used to build models, characteristics about the model (e.g, training), and the intended uses of the AI capabilities separately or as part of another system;
- Connections between and dependencies within systems, and associated potential complications;
- The selected testing methodologies and performance indicators and results for models used in the AI component (e.g., confusion matrix and thresholds for true and false positives and true and false negatives area under the curve (AUC) as metrics for performance/error); this includes how tests were done, and the simulated or real-world data used in the tests--including caveats about the assumptions of the training and testing, per type of scenarios, per the data used in testing and training;
- Required maintenance, including re-testing requirements, and technical refresh. This includes requirements for re-testing, retraining, and tuning when a system is used in a different scenario or setting (including details about definitions of scenarios and settings) or if the AI system is capable of online learning or adaptation.

3. **Infrastructure to support traceability.** Invest resources and establish policies that support the traceability of AI systems. Traceability, critical for high-stakes systems, captures key information about the system development and deployment process for relevant personnel to adequately understand the technology.[58] It includes selecting, designing, and implementing measurement tools, logging, and monitoring and applies to (1) development and testing of AI systems and components,[59] (2) operation of AI systems,[60] (3) users and their behaviors in engaging with AI systems or components,[61] and (4)

[57] For an industry example, see Timnit Gebru et al., *Datasheets for Datasets*, Microsoft (March 2018), https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/. For more on data, model and system documentation, see *Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML)*, an evolving body of work from the Partnership on AI about documentation practices at https://www.partnershiponai.org/about-ml/. See also David Thornton, *Intelligence Community Laying Foundation for AI Data Analysis*, Federal News Network (Nov. 1, 2019), https://federalnewsnetwork.com/all-news/2019/11/intelligence-community-laying-the-foundation-for-ai-data-analysis/ (documenting any caveats of re-use for both datasets and models is critical to avoid "off-label" use harms).

[58] Jonathan Mace et al., *Pivot Tracing: Dynamic Causal Monitoring for Distributed Systems*, Communications of the ACM (March 2020), https://www.cs.purdue.edu/homes/bb/cs542-20Spr/readings/others/pivot-tracing-cacm-202003.pdf [hereinafter, Mace, Pivot Tracing].

[59] Examples include logs of steps taking in problem and purpose definition, design, training and development. See e.g., Brundage, Toward Trustworthy AI Development.

[60] This includes logs of steps taken in operation which can support retrospective accident analysis. Id.

[61] Examples include logs of access and use of the system by operators, per understanding human access, oversight; nonrepudiation (e.g., cryptographic controls on access).

auditing.[62] Audits should support analyses of specific actions as well as characterizations of longer-term performance. Audits should also be done to assure that performance on tests of the system and on real-world workloads meet requirements, such as fairness asserted at specification of the system and/or established by stakeholders.[63] When a criminal investigation requires it, forensic analyses of the AI system must be supported. A recommended practice is to carefully consider how you expose APIs for audit trails and traceability infrastructure in light of the potential vulnerability to an adversary detecting how an algorithm works and conducting an attack using counter AI exploitation.

4. **Security and Robustness: Addressing Intentional and Unintentional Failures**
    - **Adversarial attacks, and use of robust ML methods**. Expand notions of adversarial attacks to include various "machine learning attacks," which may take the form of an attack through supply chain, online access, adversarial training data, or model inference attacks, including through Generative Adversarial Networks (GANS).[64]

[62] Auditing examples include real-time system health and behavior monitoring, longer-term reporting, via logging of system recommendations, classifications, or actions and why they were taken per input, internal states of the system that were important in the chain of inferences and ultimate actions, and the actions taken, and logs to assure maintenance of accountability for decision systems (e.g. signoff for a specific piece of business logic).

[63] All of the above are consistent with, and support the fulfillment of, the DOD's AI Principle, Traceable.
Documentation practices that support traceability (e.g. data sources and design procedures and documentation) are expanded upon in additional bullets throughout the Engineering Practices section. See Lopez, DOD Adopts 5 Principles ("Traceable: - The department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources and design procedures and documentation.").

[64] The approach is to simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. As the generator gets better (producing ever more credible samples) the discriminator also improves (getting ever better at discerning real samples from the generated "fake" samples). This is useful for improving discriminator performance. Given the vulnerability of deep learning models to adversarial examples (slight changes in an input that produce significantly different results in output and can be used to confound a classifier), there has been interest in using adversarial inputs in a GAN framework to train the discriminator to better distinguish adversarial inputs. There is also considerable theoretical work being done on fundamental approaches to making DL more robust to adversarial examples. This remains an important focus of research. For more on adversarial attacks, see e.g., Ian Goodfellow et al., *Generative Adversarial Networks*, Universite de Montreal (June 10, 2014), https://arxiv.org/abs/1406.2661; Ian Goodfellow et. al., *Explaining And Harnessing Adversarial Examples*, Google (Mar. 20, 2015), https://arxiv.org/pdf/1412.6572.pdf; Kevin Eykholt, et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition at 1625–1634 (2018), https://arxiv.org/abs/1707.08945; Anish Athalye, et al., *Synthesizing Robust Adversarial Examples*, International conference on machine learning (2018), https://arxiv.org/pdf/1707.07397.pdf; Kevin Eykholt, et al., *Physical Adversarial Examples for Object Detectors*, USENIX Workshop on Offensive Technologies (2018), https://arxiv.org/abs/1807.07769; Yulong Cao, et al., *Adversarial Sensor Attack on LiDAR-based Perception*

Agencies should seek latest technologies that demonstrate the ability to detect and notify operators of attacks, and also tolerate attacks. [65]

- **Follow and incorporate advances in intentional and unintentional ML failures**. Given the rapid evolution of the field of study of intentional and unintentional ML failures, national security organizations must follow and adapt to the latest knowledge about failures and proven practices for monitoring, detection, and engineering and run-time protections. Related efforts and R&D focus on developing and deploying robust AI methods.[66]
- **Adopt a security development lifecycle (SDL) for AI** systems to include a focus on potential failure modes. This includes developing and regularly refining threat models to capture and consolidate the characteristics of various attacks in a way that can shape system development to mitigate vulnerabilities.[67] A matrixed focus for developing and refining threat models is valuable. SDL should address ML development, deployment, and when ML systems are under attack.[68]

5. **Conduct red teaming** for both intentional and unintentional failure modalities. Bring together multiple perspectives to rigorously challenge AI systems, exploring the risks, limitations, and vulnerabilities in the context in which they'll be deployed.
    - To mitigate intentional failure modes – Employ methods that can make systems more resistant to adversarial attacks, work with

in Autonomous Driving, Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (2019), https://dl.acm.org/doi/10.1145/3319535.3339815; Mahmood Sharif, et al., *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016) https://dl.acm.org/doi/10.1145/2976749.2978392; Stepan Komkov & Aleksandr Petiushko, *Advhat: Real-World Adversarial Attack on Arcface Face ID System* (Aug. 23, 2019), https://arxiv.org/pdf/1908.08705.pdf. On directions with robustness, see e.g., Aleksander Madry, et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*. MIT (Sep. 4, 2019), https://arxiv.org/abs/1706.06083 [hereinafter Madry, Toward Deep Learning Models Resistant to Adversarial Attacks]; Mathias Lecuyer, et al., *Certified Robustness to Adversarial Examples with Differential Privacy*, IEEE Symposium on Security and Privacy (2019), https://arxiv.org/abs/1802.03471; Eric Wong & J. Zico Kolter, *Provable Defenses Against Adversarial Examples via the Convex Outer Adversarial Polytope*, International Conference on Machine Learning (2018), https://arxiv.org/abs/1711.00851.

[65] Madry, Towards Deep Learning Models Resistant to Adversarial Attacks.

[66] See e.g., *Id.*; Thomas Dietterich, *Steps Toward Robust Artificial Intelligence*, AI Magazine at 3-24 (Fall 2017), https://www.aaai.org/ojs/index.php/aimagazine/article/view/2756/2644; Eric Horvitz, *Reflections on Safety and Artificial Intelligence*, Safe AI: Exploratory Technical Workshop on Safety and Control for AI, White House OSTP and Carnegie Mellon University, Pittsburgh, PA (June 27, 2016), http://erichorvitz.com/OSTP-CMU_AI_Safety_framing_talk.pdf.

[67] See Andrew Marshall et al, *Threat Modeling AI/ML Systems and Dependencies* (Nov. 2010), https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml.

[68] Ram Shankar Siva Kumar et al., *Adversarial Machine Learning—Industry Perspectives*, 2020 IEEE Symposium on Security and Privacy (SP) Deep Learning and Security Workshop, (May 2020), https://arxiv.org/pdf/2002.05646.pdf.

adversarial testing tools, and deploy teams dedicated to trying to brake systems and make them violate rules for appropriate behavior.

- To mitigate unintentional failure modes - test ML systems per a thorough list of realistic conditions they are expected to operate in. When selecting third-party components, consider the impact that a security vulnerability in them could have to the security of the larger system into which they are integrated. Have an accurate inventory of third-party components and a plan to respond when new vulnerabilities are discovered.[69]
- Because of the scarcity of required expertise and experience for AI red teams, organizations should consider establishing broader enterprise-wide communities of AI red teaming capabilities that could be applied to multiple AI developments (e.g., at a DoD service or IC element level, or higher).

(4) Recommendations for Future Action

- *For documentation:* The Commission noted the urgency of a documentation strategy in its First Quarter Recommendations.[70] Future work is needed to ensure sufficient documentation by all national security departments and agencies, including the precisions noted above in this section. In the meantime, national security departments and agencies should pilot documentation approaches across the AI lifecycle to help inform such a strategy.
- *To improve traceability:* While recommended practices exist for audit trails, standards have yet to be developed.[71] Future work is needed by standard setting bodies, alongside national security departments/agencies and the broader AI community (including industry), to develop audit trail requirements per mission needs for high-stakes AI systems including safety-critical applications.
- Future R&D is needed to advance capabilities for:
  - AI security and robustness - to cultivate more robust methods that can overcome adverse conditions; advance approaches that enable assessment of types and levels of vulnerability and immunity; and to enable systems to withstand or to degrade gracefully when targeted by a deliberate attack.
  - Interpretability - to support risk assessment and better understand the efficacy of interpretability tools and possible interfaces. (Complementary

[69]See *What are the Microsoft SDL Practices?*, Microsoft (last accessed July 14, 2020), https://www.microsoft.com/en-us/securityengineering/sdl/practices.

[70] See *First Quarter Recommendations*, NSCAI (Mar. 2020) https://www.nscai.gov/reports. Ongoing efforts to share best practices for documentation among government agencies through GSA's AI Community of Practice further indicate the ongoing need and desire for common guidance.

[71] For more on current gaps in audit trail standards for AI systems, see Brundage, Toward Trustworthy AI Development at 25 ("Existing standards often define in detail the required audit trails for specific applications. For example, IEC 61508 is a basic functional safety standard required by many industries, including nuclear power. Such standards are not yet established for AI systems.").

to this R&D, standards work is needed to develop benchmarks that assess the reliability of produced model explanations.)

III. System Performance

(1) Overview

Fielding AI systems in a responsible manner includes establishing confidence that the technology will perform as intended, especially in high-stakes scenarios.[72] An AI system's performance must be assessed,[73] including assessing its capabilities and blind spots with data representative of real-world scenarios or with simulations of realistic contexts,[74] and its reliability and robustness (i.e., resilience in real-world settings—including adversarial attacks on AI components) during development and in deployment.[75] For example, a system's performance on recognition tasks can be characterized by its false positives and false negatives on a test set representative of the environment in which a system will be deployed, and test sets can be varied in realistic ways to estimate robustness. Testing protocols and requirements are essential for measuring and reporting on system performance, including reliability, during the test phase (pre-deployment) and in operational settings. (The Commission uses industry terminology 'testing' to broadly refer to what the DoD calls "Test, Evaluation, Verification, and Validation" (TEVV). This testing includes both what DoD refers to as Developmental Test and Evaluation and Operational Test and Evaluation). AI systems present new challenges to established testing protocols and requirements as they increase in complexity, particularly for operational testing. However, there are some existing methods to continuously monitor AI system performance. For example, high-fidelity performance traces and means for sensing shifts, such as distributional shifts in targeted scenarios, permit ongoing monitoring to ensure system performance does not stray outside of acceptable parameters; if inadequate performance is detected, they provide insight needed to improve and update systems.[76]

[72] This includes, for example, safety-critical scenarios or those where AI-assisted decision making would impact an individual's life or liberty.
[73] Ben Shneiderman, *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*, International Journal of Human–Computer Interaction (Mar. 23, 2020), https://doi.org/10.1080/10447318.2020.1741118 [hereinafter Shneiderman, Human Centered Artificial Intelligence: Reliable, Safe & Trustworthy].
[74] However, test protocols must acknowledge test sets may not be fully representative of real-world usage.
[75] See Brundage, Toward Trustworthy AI Development; Ece Kamar, et al., *Combining Human and Machine Intelligence in Large-Scale Crowdsourcing*, Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (June 2012), https://dl.acm.org/doi/10.5555/2343576.2343643 [hereinafter Kamar, Combining Human and Machine Intelligence in Large-Scale Crowdsourcing].
[76] For a technical paper that puts monitoring in development lifecycle context, see Amershi, Software Engineering for Machine Learning. For a good example of open source frameworks to support, see *Overview*, Prometheus, (last accessed June 18, 2020), https://prometheus.io/docs/introduction/overview/.

System performance characterization also includes assessing robustness. As noted above, this entails determining how resilient the system is in real-world settings where there may be blocking and handling of attacks and where natural real-world variation exists.[77] In addition to reliability, robustness, and security, system performance must also measure compliance with requirements derived from values such as fairness.

When evaluating system performance, it is especially important to take into account holistic, end-to-end system behavior. Emergence is the principle that entities exhibit properties which are meaningful only when attributed to the whole, not to its parts. Emergent system behavior can be viewed as a consequence of the interactions and relationships among system elements rather than the independent behavior of individual elements. It emerges from a combination of the behavior and properties of the system elements and the systems structure or allowable interactions between the elements, and may be triggered or influenced by a stimulus from the systems environment. [78]

The System Engineering Community and the National Security Community have focused on system of systems engineering for years,[79] but AI-intensive systems introduce additional opportunities and challenges for emergent performance. Given the requirement to establish and preserve justified confidence in the performance of AI systems, attention must be paid to the potential for undesired interactions and emergent performance as AI systems are composed. This composition may include pipelines where the output of one system is part of the input for another in a potentially complex and distributed ad hoc pipeline.[80] As a recent study of the software engineering challenges introduced by developing and deploying AI systems at scale notes, "AI components are more difficult to handle as distinct modules than traditional software components — models may be 'entangled' in complex ways."[81] These challenges are pronounced when the entanglement is the result of system composition and integration.

[77] Joel Lehman, *Evolutionary Computation and AI Safety: Research Problems Impeding Routine and Safe Real-world Application of Evolution* (Oct. 4, 2019), https://arxiv.org/abs/1906.10189 [hereinafter Lehman, Evolutionary Computation and AI Safety].

[78] Greg Zacharias, *Autonomous Horizons: The Way Forward*, Air University Press at 61 (Mar. 2019), https://www.airuniversity.af.edu/Portals/10/AUPress/Books/b_0155_zacharias_autonomous_horizons.pdf.

[79] Judith Dahmann & Kristen Baldwin, *Understanding the Current State of US Defense Systems of Systems and the Implications for Systems Engineering*, Presented at IEEE Systems Conference (Apr. 7-10, 2008), https://ieeexplore.ieee.org/document/4518994.

[80] D. Sculley, et al., *Machine Learning: The High Interest Credit Card of Technical Debt*, Google (2014), https://research.google/pubs/pub43146/ [hereinafter Sculley, Machine Learning: The High Interest Credit Card of Technical Debt].

[81] Amershi, Software Engineering for Machine Learning (illustrating non-monotonic error as a possible complexity result from model entanglement).

As America's AI-intensive systems may increasingly be composed (including through ad hoc opportunities to integrate systems) with allied AI-intensive systems, this becomes a topic for coordination with allies as well. Multi-agent systems are being explored and adopted in multiple domains,[82] as are swarms, fleets, and teams of autonomous systems.[83]

(2) Examples of Current Challenges

Unexpected interactions and errors commonly occur in integrated simulations and exercises as an illustration of the challenges of predicting and managing behaviors of systems composed of multiple components. Intermittent failures can transpire after composing different systems; these failures are not the result of any one component having errors, but rather are due to the interactions of the composed systems.[84]

(3) Recommendations for Adoption

Critical practices for ensuring optimal system performance are described in the following non-exhaustive list:

System Performance Recommended Practices

A. **Training and Testing: Procedures should cover key aspects of performance and appropriate performance metrics. These include:**
   1. **Standards for metrics and reporting needed to adequately achieve:**
      a. Consistency across testing and test reporting for critical areas.
      b. Testing for blinds pots as a specific failure mode of importance to some ML implementations.[85]
      c. Testing for fairness. When testing for fairness, sustained fairness assessments are needed throughout development and deployment, including assessing a system's accuracy and errors relative to one or more agreed to statistical definitions of fairness[86] and documenting deliberations

[82] Ali Dorri, et al., *Multi-Agent Systems: A Survey*, IEEE Access at 28573-28593 (Apr. 20, 2018), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8352646.

[83] Andrew Ilachinski, *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies*, CNA (Jan. 2017), https://www.cna.org/CNA_files/PDF/DRM-2017-U-014796-Final.pdf.

[84] David Sculley et al., *Hidden Technical Debt in Machine Learning Systems*, NIPS '15: Proceedings of the 28th International Conference on Neural Information Processing Systems (Dec. 2015), https://dl.acm.org/doi/10.5555/2969442.2969519.

[85] Ramya Ramakrishnan et al., *Blind Spot Detection for Safe Sim-to-Real Transfer*, Journal of Artificial Intelligence Research 67 at 191-234 (2020), https://www.jair.org/index.php/jair/article/view/11436.

[86] There is no single definition of fairness. System developers and organizations fielding applications must work with stakeholders to define fairness, and provide transparency via disclosure of assumed definitions of fairness. Definitions or assumptions about fairness and metrics for identifying fair inferences and allocations should be explicitly documented. This should be accompanied by a discussion of alternate definitions and rationales for the current choice. These elements should be documented internally as machine-learning components and larger systems are developed. This is especially important as establishing alignment on the metrics to use for assessing fairness encounters

made on the appropriate fairness metrics to use.[87] Agencies should also conduct outcome and impact analysis to detect when subtle assumptions in the system concept of operations and requirements are showing up as unexpected and undesired outcomes in the operational environment.[88]

   d. Articulation of performance standards and metrics. This includes ways to communicate to the end user the meaning/significance of performance metrics, e.g., through a probability assessment, based on sensitivity and specificity. It also requires clear documentation of system performance (across diverse environments or contexts), including information content of model output.

2. **Representativeness of the data and model for the specific context at hand.** For machine learning models, challenges exist when transferring a model to a context/setting that differs from the one for which it was trained and tested. When using classification and prediction technologies, challenges with representativeness of data used in analyses, and fairness/accuracy of inferences and recommendations made with systems leveraging that data when applied in different populations/contexts, should be considered explicitly and documented. As appropriate, robust and reliable methods can be used to enable model generalization and transfer beyond the training context.
3. **Evaluating an AI system's performance relative to current benchmarks** where possible. Benchmarks should assist in determining if an AI system's performance meets or exceeds current best performance.
4. **Evaluating aggregate performance of human-machine teams.** Consider that the current benchmark might be the current best performance of a human operator or the composed performance of the human-machine team. Where humans and machines interact, it is important to measure the aggregate performance of the team rather than the AI system alone. [89]
5. **Reliability and Robustness:** Various kinds of AI systems often demonstrate impressive performance on average, but can fail in ways that are unexpected in any specific instance. The performance potential of an AI system is often roughly determined by experiment and test, rather than by any predictive analytics. AI can have blinds spots and unknown fragilities.[90]

an added challenge when different cultural and policy norms are involved when collaborating on development and use with allies.

[87] Examples of tools available to assist in assessing and mitigating bias in systems relying on machine learning include Aequitas by the University of Chicago, Fairlearn by Microsoft, AI Fairness 360 by IBM, and PAIR and ML-fairness-gym by Google.

[88] See Microsoft's AI Fairness checklist as an example of an industry tool to support fairness assessments, Michael A. Madaio et al., *Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI*, CHI 2020 (Apr. 25-30, 2020), http://www.jennwv.com/papers/checklists.pdf [hereinafter Madaio, Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI].

[89] Kamar, Combining Human and Machine Intelligence in Large-scale Crowdsourcing.

[90] John Launchbury, *A DARPA Perspective on Artificial Intelligence*, DARPA, (last accessed June 18, 2020), https://www.darpa.mil/about-us/darpa-perspective-on-ai (noting that machine learning is "statistically impressive, but individually unreliable").

Focus on tools and techniques to carefully bound assumptions of robustness of the AI component in the larger system architecture, and provide sustained attention to characterizing the actual performance envelope for nominal and off-nominal conditions throughout development and deployment. [91]

6. **For systems of systems, testing machine-machine/multi-agent interaction**. Individual AI systems will be combined in various ways in an enterprise to accomplish broader missions beyond the scope of any single system. For example, pipelines of AI systems will exist where the output of one system serves as the input for another AI system. (The output of a track management and classifier system might be input to a target prioritization system which might in turn provide input to a weapon/target pairing tool.) Multiple relatively independent AI systems can be viewed as distinct agents interacting in the environment of the system of systems, and some of these agents will be humans in and on the loop. Industry has encountered and documented problems in building 'systems of systems' out of multiple AI systems[92] A related problem is poor backward compatibility when the performance of one model in a pipeline is enhanced and may result in degrading the overall system of system behavior.[93] These problems in composition illustrate emergent performance, as described in the conceptual overview portion of this section.

   A frequent cause of failures in composed systems is the violation of assumptions that were not previously challenged; therefore, a priority during testing should be to challenge ("stress test") interfaces and usage patterns with boundary conditions and challenges to assumptions about the operational environment and use. This is focused on both unintended violations of assumptions from system composition and also deliberate challenges to the system by adversarial attacks.

B. **Maintenance and deployment**

Given the dynamic nature of AI systems, recommended practices for maintenance are also critically important. These include:

1. **Specifying maintenance requirements** for datasets as well as for systems, given that their performance can degrade over time.[94]
2. **Continuously monitoring and evaluating AI system performance**, including the use of high-fidelity traces to determine continuously if a system

[91] Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy.

[92] One example is "Hidden Feedback Loops", where systems that learn from external world behavior may also shape the behavior they are monitoring. See Sculley, Machine Learning: The High Interest Credit Card of Technical Debt. See also Cynthia Dwork, et al., *Individual Fairness in Pipelines*, (apr. 12, 2020), https://arxiv.org/abs/2004.05167; Megha Srivastava, et al., *An Empirical Analysis of Backward Compatibility in Machine Learning Systems*, KDD '20 (forthcoming, August 2020) [hereinafter Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems].

[93] Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems.

[94] *Artificial Intelligence (AI) Playbook for the U.S. Federal Government*, Artificial Intelligence Working Group, ACT-IAC Emerging Technology Community of Interest, (January 22, 2020), https://www.actiac.org/act-iac-white-paper-artificial-intelligence-playbook.

is going outside of acceptable parameters (including operational performance measures and established constraints for fairness and core values), both during pre-deployment and operation.[95] This includes measuring system performance per acceptable parameters in terms of both reliability and values.[96] It also includes assessing statistical results for performance over time, for example, to detect emergent bias or anomalies.[97]

3. **Iterative and sustained testing and validation**. Be wary that training and testing that provide characteristics on capabilities might not transfer or generalize to specific settings of usage (for example lighting conditions in some applications may be very different for scene interpretation); thus, testing and validation may need to be done recurrently, and at strategic intervention points, but especially for new deployments and classes of task.[98]
4. **Monitoring and mitigating emergent behavior**. There will be instances where systems are composed in ways not anticipated by the developers (e.g., opportunistic integration with an ally's system). These use cases clearly can't be adequately addressed at development time; some aspects of confidence in the composition must be shifted to monitoring the actual performance of the composed system and its components. For emergent performance concerns when AI systems are composed, there are

[95] Beyond accuracy, high-fidelity traces capture other parameters of interest/musts, including fairness, fragility (e.g. whether a system degrades gracefully versus unexpectedly fails), security/attack resilience, and privacy leakage. Often instrumentation results from execution are treated as time-series data and can be analyzed by a variety of anomaly detection techniques to identify unexpected or changing characteristics of system performance. See Meir Toledano et al., *Real-Time Anomaly Detection System for Time Series at Scale*, KDD 2017: Workshop on Anomaly Detection in Finance (2017), http://proceedings.mlr.press/v71/toledano18a/toledano18a.pdf. DOD recently updated its acquisition processes to improve "the ability to deliver warfighting capability at the speed of relevance" See *DoD 5000 Series Acquisition Policy Transformation Handbook*, Department of Defense (Jan. 15, 2020), https://www.acq.osd.mil/ae/assets/docs/DoD%205000%20Series%20Handbook%20(15Jan2020).pdf. These include revised policies for acquiring software-intensive systems and components. Relevant here, program managers are now required to "ensure that software teams use iterative and incremental software development methodologies," and use modern technologies "to achieve automated testing, continuous integration and continuous delivery of user capabilities, frequent user feedback/engagement (at every iteration if possible), security and authorization processes, and continuous runtime monitoring of operational software" Ellen Lord, *Software Acquisition Pathway Interim Policy and Procedures*, Memorandum from the Undersecretary of Defense, to Joint Chiefs of Staff and Department of Defense Staff (Jan. 3, 2020), https://www.acq.osd.mil/ae/assets/docs/USA002825-19%20Signed%20Memo%20(Software).pdf. See also Ori Cohen, *Monitor! Stop Being A Blind Data-Scientist* (Oct. 8, 2019), https://towardsdatascience.com/monitor-stop-being-a-blind-data-scientist-ac915286075f; Mace, Pivot Tracing.

[96] Values parameters could include pre-determined thresholds for acceptable false positive or false negative rates for fairness, or parameters set regarding data or model leakage in the context of privacy.

[97] Lehman, Evolutionary Computation and AI Safety.

[98] Eric Breck, et al., *The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction*, 2017 IEEE International Conference on Big Data, (Dec. 11-14, 2017), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8258038&tag=1.

advances in runtime assurance/verification[99] and feature interaction management[100] that can be adapted.

(4) Recommendations for Future Action

- Future R&D is needed to advance capabilities for:
    - Testing, Evaluation, Verification, and Validation (TEVV) of AI systems - to develop a better understanding of how to conduct TEVV and build checks and balances into an AI system. Includes complex system testing - to increase our understanding of and ability to have confidence in emergent performance of composed AI systems. Improved methods are needed to understand, predict, and control systems-of-systems so that when AI systems interact with each other, their interaction does not lead to unexpected negative outcomes.
    - Multi-agent scenario understanding - to advance the understanding of interacting AI systems, including the application of game theory to varied and complex scenarios, and interactions between cohorts composed of a mixture of humans and AI technologies.
- Basic definitional work has been ongoing for years on how to characterize key properties such as fairness and explainability. Progress on a common understanding of the concepts and requirements is critical for progress in widely used metrics for performance.
- Significant work is needed to establish what appropriate metrics should be to assess system performance across attributes for responsible AI and across profiles for particular applications/contexts. (Such attributes, for example, include fairness, interpretability, reliability and robustness.)
- International collaboration and cooperation is needed to:
    - Align on how to test and verify AI system reliability and performance along shared values (such as fairness and privacy). Establishing how to test systems will include measures of performance based on common standards, and may have implications for the types of traceability that will need to be incorporated into system design and development. Such collaboration on common testing for reliability and adherence to

[99] Shuvendu Lahiri, et al., *Runtime Verification*, 17th International Conference on Runtime Verification (Sept. 13-16, 2017), https://link.springer.com/book/10.1007/978-3-319-67531-2; Christian Colombo ,et al., *Runtime Verification*, 18th International Conference on Runtime Verification (Nov. 10-13, 2018), https://link.springer.com/book/10.1007/978-3-030-03769-7; Sanjit A. Seshia, *Compositional Verification without Compositional Specification for Learning-Based Systems*, UC Berkeley (Nov. 26, 2017), https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-164.pdf.

[100] Larissa Rocha Soares, et al., *Feature Interaction in Software Product Line Engineering: A Systematic Mapping Study*, Information and Software Technology at 44-58 (June 2018), https://www.sciencedirect.com/science/article/abs/pii/S0950584917302690; Seregy Kolesnikov, *Feature Interactions in Configurable Software Systems*, Universität Passau (Aug. 2019), https://www.researchgate.net/publication/334926566_Feature_Interactions_in_Configurable_Software_Systems; Bryan Muscedere, et al., *Detecting Feature-Interaction Symptoms in Automotive Software using Lightweight Analysis*, 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering at 175-185 (2019), https://ieeexplore.ieee.org/document/8668042.

> values will be critical among allies and partners to enable interoperability and trust. Additionally, these efforts could potentially include dialogues between the United States and strategic competitors regarding establishing common standards of AI safety and reliability testing in order to reduce the chances of inadvertent escalation.[101]

IV. Human-AI Interaction

(1) Overview

Responsible AI development and fielding requires striking the right balance of leveraging human and AI reasoning, recommendation, and decision-making processes. Ultimately, all AI systems will have some degree of human-AI interaction as they will all be developed to support humans. In some settings, the best outcomes will be achieved when AI is designed to augment human intellect, or to support human-AI collaboration more generally. In other settings, however, time-criticality and the nature of tasks may make some aspects of human-AI interaction difficult or suboptimal.[102] Where the human role is critical in real-time decisions because it is more appropriate, valuable, or designated as such by our values, AI should be intentionally designed to effectively augment and support human understanding, decision making, and intellect. Sustained attention must be focused on optimizing the desired human-machine interaction throughout the AI system lifecycle. It is important to think through the use criteria that are most relevant depending on the model. Models are different for human-assisted AI decision-making, AI-assisted human decision-making, pure AI decision-making, and AI-assisted machine decision-making.

(2) Examples of Current Challenges

There is an opportunity to develop AI systems to complement and augment human understanding, decision making, and capabilities. Decisions about developing and fielding AI systems aimed at specific domains or scenarios should consider the relative strengths of AI capabilities and human intellect across expected distributions of tasks, considering AI system maturity or capability and how people and machines might coordinate.

Designs and methods for human-AI interaction can be employed to enhance human-

[101] For research regarding common interests in ensuring safety-critical systems work as intended (e.g. in a reliable manner) to avoid destabilization/escalatory dynamics, see Andrew Imbrie & Elsa Kania, *AI Safety, Security, and Stability Among Great Powers Options, Challenges, and Lessons Learned for Pragmatic Engagement*, CSET, (Dec. 2019), https://cset.georgetown.edu/wp-content/uploads/AI-Safety-Security-and-Stability-Among-the-Great-Powers.pdf.

[102] The need for striking the right balance of human involvement in situations of time criticality is not unique to AI. For instance, DoD systems dating back to the 80s have been designed to react to airborne threats at speeds faster than a human would be capable of. See *MK 15 - Phalanx Close-In Weapons System (CIWS)*, U.S. Navy (last accessed June 18, 2020), https://www.public.navy.mil/surfor/Pages/Phalanx-CIWS.aspx.

AI teaming.[103] Methods in support of effective human-AI interaction can help AI systems to understand when and how to engage humans for assistance, when AI systems should take initiative to assist human operators, and, more generally, how to support the creation of effective human-AI teams. In engaging with end users, it may be important for AI systems to infer and share with end users well-calibrated levels of confidence about their inferences, so as to provide human operators with an ability to weigh the importance of machine output or pause to consider details behind a recommendation more carefully. Methods, representations, and machinery can be employed to provide insight about AI inferences, including the use of interpretable machine learning.[104] Research directions include developing and fielding machinery aimed at reasoning about human strengths and weaknesses, such as recognizing and responding to the potential for costly human biases of judgment and decision making in specific settings.[105] Other work centers on mechanisms that consider the ideal mix of initiatives, including when and how to rely on human expertise versus on AI inferences.[106] As part of effective teaming, AI systems can be endowed with the ability to detect the focus of attention, workload, and interruptability of human operators and consider these inferences in decisions about when and how to engage with the operators.[107] Directions of effort include developing mechanisms for identifying the most relevant information or inferences to provide end users of different skills in different settings.[108] Consideration must be given to the prospect introducing bias, including potential biases that may arise because of the configuration and sequencing of rendered data. For example, IC research[109] shows that confirmation bias can be triggered by the order in which information is

[103] Saleema Amershi, et al., *Guidelines for Human-AI Interaction*, Proceedings of the CHI Conference on Human Factors in Computing Systems (2019), https://dl.acm.org/doi/10.1145/3290605.3300233

[104] Rich Caruana, et al., Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Semantic Scholar (Aug. 2015), https://www.semanticscholar.org/paper/Intelligible-Models-for-HealthCare%3A-Predicting-Risk-Caruana-Lou/cb030975a3dbcdf52a01cbd1c140711332313e13.

[105] Eric Horvitz, *Reflections on Challenges and Promises of Mixed-Initiative Interaction*, AAAI Magazine 28 Special Issue on Mixed-Initiative Assistants (2007), http://erichorvitz.com/mixed_initiative_reflections.pdf.

[106] Eric Horvitz, *Principles of Mixed-Initiative User Interfaces*, Proceedings of CHI '99 ACM SIGCHI Conference on Human Factors in Computing Systems (May 1999), https://dl.acm.org/doi/10.1145/302979.303030; Kamar, Combining Human and Machine Intelligence in Large-scale Crowdsourcing.

[107] Eric Horvitz, et al., *Models of Attention in Computing and Communications: From Principles to Applications*, Communications of the ACM 46(3) at 52-59 (Mar. 2003), https://cacm.acm.org/magazines/2003/3/6879-models-of-attention-in-computing-and-communication/fulltext.

[108] Eric Horvitz & Matthew Barry, *Display of Information for Time-Critical Decision Making*, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (Aug. 1995), https://arxiv.org/pdf/1302.4959.pdf.

[109] There has been considerable research in the IC on the challenges of confirmation bias for analysts. Some experiments demonstrated a strong effect that the sequence in which information is presented alone can shape analyst interpretations and hypotheses. Brant Cheikes, et al., *Confirmation Bias in Complex Analyses*, MITRE (Oct. 2004), https://www.mitre.org/sites/default/files/pdf/04_0985.pdf. This highlights the care that is required when designing the human machine teaming when complex, critical, and potentially ambiguous information is presented to analysts and decision makers.

displayed, and this order can consequently impact or sway intel analyst decisions. Careful design and study can help to identify and mitigate such bias.

(3) Recommendations for Adoption

Critical practices to ensure optimal human-AI interaction are described in the non-exhaustive list below. These recommended practices span the entire AI lifecycle.

Human-AI Interaction Recommended Practices

A. **Identification of functions of human in design, engineering, and fielding of AI**
   1. **Define functions and responsibilities of human operators and assign them to specific individuals**. Functions will vary for each domain and each project within a domain; they should be periodically revisited as model maturity and human expertise evolve over time.
   2. **Given the nature of the mission and current competencies of AI, policies should define the tasks of humans across the AI lifecycle,** noting needs for feedback loops, including opportunities for oversight.
   3. **Enable feedback and oversight to ensure that systems operate as they should** - algorithmic accountability means that there is a governance structure in place to correct grievances if systems fail.

B. **Explicit support of human-AI interaction and collaboration**
   1. **Human-AI design guidelines**. AI systems designs should take into consideration the defined tasks of humans in human-AI collaborations in different scenarios; ensure the mix of human-machine actions in the aggregate is consistent with the intended behavior, and accounting for the ways that human and machine behavior can co-evolve;[110] and also avoid automation bias and unjustified reliance on humans in the loop as failsafe mechanisms. Allow for auditing of the human-AI pair, not only the AI in isolation, which could be a secondary expert examining a subset of cases. Designs should be transparent (e.g., about why and how a system did what it did, system updates, or new capabilities) so that there is an understanding the AI is working day-to-day and to allow for an audit trail if things go wrong .[111] Based on context and mission need, designs should ensure usability of AI systems by AI experts, domain experts, and novices, as appropriate.[112] Both transparency and usability will depend on the audience.

[110] Patricia L. McDermott et al.,, *Human-machine Teaming Systems Engineering Guide*, MITRE (Dec. 2018), https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide; Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy.

[111] For additional examples, see *Guidelines for Human AI Interaction*, Microsoft (June 4, 2019), https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/.

[112] Systems are sometimes designed with the assumption of a human in the loop as the failsafe or interlock, but humans often defer to computer generated results and get in the habit of confirming machine results without scrutiny.

2. **Algorithms and functions in support of interpretability and explanation.** Algorithms and functions that provide individuals with task-relevant knowledge and understanding need to take into consideration that key factors in an AI system's inferences and actions can be understood differently by various audiences. These audiences span real-time operators who need to understand inferences and recommendations for decision support, engineers and data scientists involved in developing and debugging systems, and other stakeholders including those involved in oversight. Interpretability and explainability exists in degrees; what's needed in terms of explainability will depend on who is receiving the explanation, what the context is, and the amount of time available to deliver and process this explanation. In this regard, interpretability intersects with traceability, audit, and documentation practices.
3. **Designs that provide cues to the human operator(s) about the level of confidence the system has in the results or behaviors of the system**.[113] AI system designs should appropriately convey uncertainty and error bounding. For instance, a user interface should convey system self-assessment of confidence alerts when the operational environment is significantly different from the environment the system was trained for, and indicate internal inconsistencies that call for caution.
4. **Policies for machine-human initiative and handoff.** Policies, and aspects of human computer interaction, system interface, and operational design, should define when and how information or tasks should be handed off from a machine to a human operator and vice versa. Include checks to continually evaluate whether distribution of tasks is working. Special attention should be given to the fact that humans may freeze during an unexpected handoff due to the processing time the brain needs, potential distractions, or the condition during which the handoff occurs. The same may be true with an AI system which may not fully understand the human's intent during the handoff and may consequently make unexpected actions.
5. **Leveraging traceability to assist with system development and understanding**. Traceability processes must include audit logs or other traceability mechanisms to retroactively understand if something went wrong, and why, in order to improve systems and their use in the future and for redress. Infrastructure and instrumentation[114] can also help assess humans, systems, and environments to gauge the impact of AI at all levels of system maturity; and to measure the effectiveness and performance for hybrid human-AI systems in a mission context.

[113] When systems report confidence in probabilities of correctness, these should be well calibrated. At the same time, it is important to acknowledge that there are limits to the confidence that can be assigned to a system estimate of correctness.

[114] Infrastructure includes tools (hardware and software) in the test environment that support monitoring system performance (such as the timing of exchanges among systems, or the ability to generate test data). Instrumentation refers to the presence of monitoring and additional interfaces to provide insight into a specific system under test.

6. **Training**. Train and educate individuals responsible for AI development and fielding, including human operators, decision makers, and procurement officers. Training should include experiences with use of systems in realistic situations. Beyond training in the specifics of the system and application, operators of systems with AI components, especially systems that perform classification or pattern recognition, should receive education that includes fundamentals of AI and data science, including coverage of key descriptors of performance, including rates of false negatives and false positives, precision and recall, and sensitivity and specificity.

   Periodic certification and refresh. In addition to initial programs of training, operators should receive ongoing refresher trainings. Beyond being scheduled periodically, refresher trainings are appropriate when systems are deployed in new settings and unfamiliar scenarios. Refresh on training is also needed when predictive models are revised with new or additional data as the performance of systems may shift with such updates introducing behaviors that are unfamiliar to human operators.[115]

(4) Recommendations for Future Action

- Future R&D is needed to advance capabilities for:
  - Enhanced human-AI interaction -
    - To progress the ability of AI technologies to perceive and understand the meaning of human communication, including spoken speech, written text, and gestures. This research should account for varying languages and cultures, with special attention to diversity given that AI typically performs worse in cases with gender and racial minorities.
    - To improve human-machine teaming. This should include disciplines and technologies centered on decision sciences, control theory, psychology, economics (human aspects and incentives), and human factors engineering, such as human-AI interfaces**,** to enhance situational awareness and make it easier for users to do their work. Human-AI interaction and the mechanisms and interfaces that support such interactions, including richer human-AI *collaborations*, will depend upon mission needs and appropriate degrees of autonomy versus human oversight and control. R&D for human-machine teaming should also focus on helping systems understand human blind spots and biases, and optimizing factors such as human attention, human workload, ideal mixing of human and machine initiatives, and passing control between the human and machine. For effective passing of control, and to

[115] Gagan Bansal et al., *Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff*, AAAI (Jul. 2019), https://www.aaai.org/ojs/index.php/AAAI/article/view/4087.

have effective and trusted teaming, R&D should further enable humans and machines to better understand intent and context of handoff.

- Ongoing work is needed to train the workforce that will interact with, collaborate with, and be supported by AI systems. In its First Quarter Recommendations, the Commission provided recommendations for such training.[116]
    - **Workforce training**. A complementary best practice for Human-AI Interaction is training the workforce to understand tools they're using; as AI gets democratized, it will also get misused. For probabilistic systems, concepts and ideas that are important in system operation should be understood; for operators this includes understanding concepts such as precision, recall, sensitivity and specificity, and ensuring operators know how to interpret the confidence in inferences that well-calibrated systems convey.

V. Accountability and Governance

(1) Overview

National security departments and agencies must specify who will be held accountable for both specific system outcomes and general system maintenance and auditing, in what way, and for what purpose. Government must address the difficulties in preserving human accountability, including for end users, developers, testers, and the organizations employing AI systems. End users and those ultimately affected by the actions of an AI system should be offered the opportunity to appeal an AI system's determinations. And, finally, accountability and appellate processes must exist not only for AI decisions, but also for AI system inferences, recommendations, and actions.

(2) Examples of Current Challenges

Overseeing entities must have the technological capacity to understand what in the AI system caused the contentious outcome. For example, if a soldier uses an AI-enabled weapon and the result violates international law of war standards, an investigating body or military tribunal should be able to re-create what happened through auditing trails and other documentation. Without policies requiring such technology and the enforcement of those policies, proper accountability would be elusive if not impossible. Moreover, auditing trails and documentation will prove critical as courts begin to grapple with whether AI system's determinations reach the requisite standards to be admitted as evidence.[117] Building the traceability infrastructure to permit auditing (as described in the Engineering Practices section)

[116] See *First Quarter Recommendations*, NSCAI (Mar. 2020), https://www.nscai.gov/reports.

[117] For more on the difficulties of admitting ML evidence, see Patrick Nutter, *Machine Learning Evidence: Admissibility and Weight*, University of Pennsylvania Law (Feb. 2019), https://scholarship.law.upenn.edu/jcl/vol21/iss3/8/.

will increase the costs of building AI systems and take significant work -- a necessary investment given our commitment to accountability, discoverability, and legal compliance.

(3) Recommendations for Adoption

Critical accountability and governance practices are identified in the non-exhaustive list below.

Accountability and Governance Recommended Practices

1. **Identify responsible actors.** Determine and document the human beings accountable for a specific AI system or any given part of an AI system and the processes involved with it. This includes identifying persons responsible for the operation of an AI system including the system's inferences, recommendations, and actions during usage, as well as the enforcement of policies for using a system. Determine and document the mechanism/structure for holding such actors accountable and to whom should that mechanism/structure be disclosed to ensure proper oversight.
2. **Adopt technology to strengthen accountability processes and goals**. Document the chains of custody and command involved in developing and fielding AI systems. This will allow the government to know who was responsible at which point in time. Improving traceability and auditability capabilities will allow agencies to better track a system's performance and outcomes. [118]
3. **Adopt policies to strengthen accountability.** Identify or, if lacking, establish policies that allow individuals to raise concerns about irresponsible AI, e.g. via an ombudsman. Agencies should institute specific oversight and enforcement practices, including: auditing and reporting requirements, a mechanism that would allow thorough review of the most sensitive/high-risk AI systems to ensure auditability and compliance with other responsible use and fielding requirements, an appealable process for those who have been found at fault of developing or using AI irresponsibly, and grievance processes for those affected by the actions of AI systems. Agencies should leverage best practices from academia and industry for conducting internal audits and assessments,[119] while also acknowledging the benefits offered by external audits.[120]

[118] See Raji, Closing the AI Accountability Gap.

[119] See Raji, Closing the AI Accountability Gap ("In this paper, we present internal algorithmic audits as a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles"); see also Madaio, Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI.

[120] For more on the benefits of external audits, see Brundage, Toward Trustworthy AI Development. For an agency example, see Aaron Boyd, *CBP Is Upgrading to a New Facial Recognition Algorithm in March*, Nextgov.com (Feb. 7, 2020), https://www.nextgov.com/emerging-tech/2020/02/cbp-upgrading-

4. **External oversight support**. Remain responsive and facilitate Congressional oversight through documentation processes and other policy decisions.[121] For instance, supporting traceability and specifically documentation to audit trails, will allow for external oversight.[122] Internal self-assessment alone might prove to be inadequate in all scenarios.[123] Congress can provide a key oversight function throughout the AI lifecycle, asking critical questions of agency leadership and those responsible for AI systems.

(4) Recommendations for Future Action

- Currently no external oversight mechanism exists specific to AI in national security. Notwithstanding the important work of Inspectors General in conducting internal oversight, open questions remain as to how to complement current practices and structures.

Appendix A — DoD AI Principles Alignment Table

NSCAI staff developed the below table to illustrate how U.S. government AI ethics principles, like those recently issued by the DoD, can be operationalized through NSCAI's Key Considerations for Responsible Development and Fielding of AI (See Appendix A-1 and A-2). Other Federal agencies and departments can use this table to visualize how NSCAI's recommended practices align with their own AI principles, or as guidance in the absence of internal AI ethics principles. In the table below, an "X" indicates that the NSCAI recommended practice on the left operationalizes the DoD principle at the top. As the table shows, every NSCAI recommended practice implements one or more DOD AI ethics principles. And every DoD AI ethics principle has at least one recommended practice that implements the principle.

new-facial-recognition-algorithm-march/162959/ (highlighting a NIST algorithmic assessment on behalf of U.S. Customs and Border Protection).

[121] Maranke Wieringa, *What to Account for When Accounting for Algorithms*, Proceedings of the 2020 ACM FAT Conference, (Jan. 2020), https://dl.acm.org/doi/10.1145/3351095.3372833.

[122] Raji, Closing the AI Accountability Gap.

[123] Brundage, Toward Trustworthy AI Development.

	NSCAI Recommended Practices:	**DOD PRINCIPLES OF AI ETHICS**				
		Responsible	**Equitable**	**Traceable**	**Reliable**	**Governable**
Core Values	**A1 -** Employ technologies and operational policies for privacy, fairness, inclusion, human rights, and law of armed conflict	X	X			X
	B1 - Consider and document value considerations based on how tradeoffs with accuracy are handled	X	X	X	X	
	B2 - Consider and document value considerations in systems that rely on representations of objective or utility functions	X	X	X	X	
	B3 - Conduct documentation, reviews, and set limits on disallowed outcomes	X	X	X	X	X
Engineering	**1 -** Concept of operations development, and design and requirements definition and analysis	X	X	X	X	X
	2 - Documentation of the AI lifecycle			X		
	3 - Infrastructure to support traceability, including auditability and forensics		X	X		
	4 - Security and robustness: addressing intentional and unintentional failures				X	X
	5 - Conduct red-teaming				X	
System Performance	**A1 -** Standards for metrics & reporting	X	X	X	X	
	A2 - Representativeness of data and model for the specific context at hand	X	X	X	X	
	A3 - Evaluating an AI system's performance relative to current benchmarks			X		X
	A4 - Evaluating aggregate performance of human-machine teams	X				
	A5 - Reliability and robustness	X		X	X	
	A6 - For systems of systems, testing machine-machine/multi-agent interaction	X			X	
	B1 - Specifying maintenance requirements	X	X	X	X	
	B2 - Continuously monitoring and evaluating AI system performance	X	X	X	X	X
	B3 - Iterative and sustained testing and validation	X			X	X
	B4 - Monitoring and mitigating emergent behavior	X			X	X
Human-AI Interaction	**A1 -** Define functions and responsibilities of human operators and assign them to specific individuals	X		X		
	A2 - Policies should define the tasks of humans across the AI lifecycle	X				
	A3 - Enable feedback and oversight to ensure that systems operate as they should	X			X	
	B1 - Human-AI design guidelines	X	X	X		X
	B2 - Algorithms and functions in support of interpretability and explanation	X		X		X
	B3 - Designs that provide cues to human operator(s) about the confidence a system has in its results or behaviors	X		X		X
	B4 - Policies for machine-human handoff	X		X		X
	B5 - Leveraging traceability to assist with system development and understanding	X		X	X	X
	B6 - Training	X	X	X	X	X
Accountability/ Governance	**1 -** Identify responsible actors	X		X		X
	2 - Adopt technology to strengthen accountability processes and goals	X		X		X
	3 - Adopt policies to strengthen accountability	X		X		
	4 - External oversight support			X		