Key Considerations for Responsible Development & Fielding of Artificial Intelligence

Abridged Version

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 Council 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.

The content herein is an abridged version of the content included here as well as in Tab 6 of the Commission's July 2020 report.

Introduction

The Commission acknowledges the efforts undertaken to date to establish ethics guidelines for AI systems.[1] While some national security agencies have adopted,[2] or are in the process of adopting, AI principles,[3] other 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. 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.

This Commission is identifying a set of challenges and making recommendations on directions with responsibly developing and fielding AI systems, and for pinpointing the concrete actions that should be adopted across the government to help overcome these challenges. Collectively, they form a paradigm for aligning AI system development and AI system behavior to goals and values. The first section, *Aligning Systems and Uses with American Values and the Rule of Law*, provides guidance specific to implementing systems that abide by American values, most of which are shared by democratic nations. The section also 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 further outline practices needed to develop and field systems that are trustworthy, understandable, reliable, and robust.

Recommended practices span multiple phases of the *AI lifecycle*, and establish a baseline for the responsible development and fielding of AI technologies. 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.'

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 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 through contractors) should be subjected to the same rigorous standards and recommended 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.

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 examples 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 best practices do not exist or 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.

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). These values are codified in the U.S. Constitution and the U.S. Code.[4] Our values also are found in international treaties that the United States has ratified that affirm our commitments to human rights and human dignity.[5] Within America's national security departments, our commitment to protecting and upholding privacy and civil liberties is further embedded in the policies and programs of the Intelligence Community,[6] the Department of Homeland Security,[7] the Department of Defense (DoD),[8] and oversight entities.[9] 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.[10] Other values are reflected in treaties, rules, and policies such as the Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment;[11] the DoD's Rules of Engagement;[12] and the DoD's Directive 3000.09.[13] While not an exhaustive list of U.S. values, the paradigm of considerations and recommended practices for AI that we introduce resonate with these values as they have been acknowledged 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.

Our 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 our departmental policies. In short, American values must inform the way we develop and field AI systems, and the way our AI systems behave in the world.

In the more comprehensive document (Appendix A-2), we provide additional details and references for technical implementers and note where recommendations would support the fulfillment of the high-level AI principles that have been adopted by the Secretary of Defense.

(2) Examples of Current Challenges

Machine learning (ML) techniques can assist DoD agencies with large-scale data analyses to support and enhance decision making about personnel. As an example, the Proposed New Disability Construct (PNDC) seeks to leverage data analyses to identify service members on the verge of ineligibility due to concerns with their readiness. Other potential analyses can support personnel evaluations, including analyzing 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 areas like criminal justice, recruiting and hiring, and face recognition.[14] Attention should be paid to challenges with decision support systems to avoid harmful disparate impact.[15] Likewise, factors weighed in performance evaluations and promotions must be carefully considered to avoid inadvertently reinforcing existing biases through ML-assisted decisions.[16]

(3) Recommendations for Adoption

A. **Developing uses and building systems that behave in accordance with American values and the rule of law.** To implement core American values, it is important to:
   1. **Employ technologies and operational policies that align with privacy preservation, fairness, inclusion, human rights, and the law of armed conflict (LOAC).** Technologies and policies throughout the AI lifecycle should support achieving these goals; they should ensure that AI uses and systems are consistent with these values and mitigate the risk that AI system uses/outcomes will violate these values.

B. **Representing Objectives and Trade-offs**. 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.[17] Recommended Practices for Representing Objectives

and Trade-offs include the following:

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.**

(4) Recommendations for Future Action

Future R&D. 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.[18] 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), the use of anonymity techniques and privacy-preserving technologies, and uses of related technologies such as multiparty computation (to allow collaboration on the pooling of data from multiple organizations without sharing datasets). Additionally, we need to understand the compatibility of data usage policies and privacy preserving approaches with regulatory approaches such as the European Union's General Data Protection Regulation (GDPR).

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[19]*; 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[20]); *interpretable* (to understand the value and accuracy of system output[21])*, and reliable* (to perform in the intended manner within the intended domain of use[22]). There are no broadly directed best practices or standards to guide organizations in the building of AI systems that are consistent with designated AI principles, but candidate approaches, minimal standards, and engineering proven practices are available.[23]

Additionally, several properties of the methods and models used in ML (e.g., data-centric methods) 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.[24] Recent efforts integrate adversarial attacks[25] and unintended faults throughout the lifecycle[26] into a single framework that recognizes intentional and unintentional failure modes.[27]

Intentional failures are the result of malicious actors explicitly attacking some aspect of (AI) system behavior. Taxonomies on malicious attacks explain the rapidly developing Adversarial Machine Learning *(*AML*)* landscape. Attacks span ML training and testing and each have associated defenses.[28] Categories of intentional failures introduced by adversaries include training *data poisoning* attacks (contaminating training data), *model inversion* (recovering secret features used in the model through careful queries), and ML *supply chain attacks* (comprising the ML model as it is being downloaded for use).[29] National security uses of AI will be the subject of sustained adversarial efforts; AI developed for this community must remain current with a rapidly developing understanding of the nature of vulnerabilities to attacks as these attacks grow in sophistication. Technical and process advances that contribute to reducing vulnerability and to detecting and alerting about attacks must also be monitored routinely.

Unintentional failures can be introduced at any point in the AI development and deployment lifecycle. In addition to faults that can be inadvertently introduced into any software development effort, distinct additional failure modes can be introduced for machine learning systems.

Examples of unintentional AI failures include *reward hacking* (when AI systems act counter to the intent of the programmed rules because of a mismatch between stated reward and real reward) and *distributional shifts* (when a system is tested in one kind of environment, but is unable to adapt to changes in other kinds of environment).[30] 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.[31] As AI systems that are separately developed and tested are composed and interact with other AI systems (within one's own services, forces, agencies, and between US systems and those of allies, adversaries, and potential adversaries), additional unintentional failures can occur.[32]

(2) Examples of Current Challenges

To make high-stakes decisions, and often in safety-critical contexts, the DoD and Intelligence Community (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 IC,[33] including concerns with detecting adversarial attacks such as data poisoning.

(3) Recommendations for Adoption

Critical engineering practices needed to operationalize AI principles (such as 'traceable' and 'reliable'[34]) 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 and potential functions that can be performed by an AI technology. Assess general feasibility of specific candidate AI technologies, based on analyses of use cases and scenario development. This includes broad stakeholder engagement and hazard analysis with multi-disciplinary experts that ask key questions about potential disparate impact and document the process undertaken to ensure fairness and lack of unwanted bias in the ML application.[35] The feasibility of meeting these requirements may trigger a review of whether and where it is appropriate to use AI in the system being proposed.
    - **Risk assessment**. Trade-offs and risks, including a system's potential societal impact, should be discussed with a diverse, interdisciplinary group. Risk assessment questions should be asked about critical areas relevant to the national security context, including privacy and civil liberties, LOAC, human rights,[36] system security, and the risks of a new technology being leaked, stolen, or weaponized.[37]
2. **Documentation of the AI lifecycle:** Whether building and fielding an AI system or "infusing AI" into a preexisting system, require documentation in certain areas.[38] These include the data used in ML and origin of the data;[39] algorithm(s) used to build models, model characteristics, and intended uses of the AI capabilities; connections between and dependencies within systems, and associated potential complications; the selected testing methodologies, performance indicators, and results for models used in the AI component; and required maintenance (including re-testing requirements) and technical refresh (including for when a system is used in a different scenario/setting or if the AI system is capable of online learning or adaptation).
3. **Infrastructure for traceability.** Invest resources and establish policies that support the traceability of AI systems. Traceability captures key information about the system development and deployment process for relevant personnel to adequately understand the technology.[40] Audits should support analyses of specific actions and characterizations of longer-term performance, and assure that performance on tests of the system and on real-world workloads meet requirements.
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,"[41] and seek latest technologies that demonstrate the ability to detect and notify operators of attacks, and also tolerate attacks.[42]

- **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.[43]
- **Adopt a security development lifecycle (SDL) for AI** systems focused on potential failure modes. This includes developing and regularly refining threat models to capture and characteristics of various attacks, establish a matrixed focus for developing and refining threat models, and ensuring SDL addresses ML development, deployment, and when ML systems are under attack.[44]

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 (i.e., red teaming).
   - To mitigate intentional failure modes - Use methods to make systems more resistant to adversarial attacks, work with adversarial testing tools, and deploy teams dedicated to trying to brake systems and make them violate rules for appropriate behavior. [45]
   - 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.[46]
   - 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

- **Documentation strategy.** As noted in our First Quarter Recommendations, a common documentation strategy is needed to ensure sufficient documentation by all national security departments and agencies.[47] In the meantime, agencies should pilot documentation approaches across the AI lifecycle to help inform such a strategy.
- **Standards**. To improve traceability, future work is needed by standard setting bodies, alongside national security departments/agencies and the broader AI community, to develop audit trail requirements per mission needs for high-stakes AI systems including safety-critical applications.
- **Future R&D.** R&D is needed to advance capabilities for cultivating more robust methods that can overcome adverse conditions; to 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. R&D is also needed to advance capabilities to support risk assessment; to better understand the efficacy of interpretability tools and possible interfaces; and 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. An AI system's performance must be assessed,[48] including assessing its capabilities and blind spots with data representative of real-world scenarios or with simulations of realistic contexts,[49] and its reliability, robustness (i.e., resilience in real-world settings—including adversarial attacks on AI components), and security during development and deployment.[50] System performance must also measure compliance with requirements derived from values such as fairness.

Testing protocols and requirements are essential for measuring and reporting on system performance. (Here, 'testing' broadly refers 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, existing methods like high-fidelity performance traces and means for sensing shifts, such as distributional shifts in targeted scenarios, allow for the continuous monitoring of an AI system's performance.

When evaluating system performance, it is especially important to take into account holistic, end-to-end system behavior—the consequence of the interactions and relationships among system elements rather than the independent behavior of individual elements. While system engineering and national security communities have focused on system of systems engineering for years, specific attention must be paid to undesired interactions and emergent performance in AI systems. 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.[51] A related problem is encountered when the performance of one model in a pipeline changes, degrading the overall pipeline behavior.[52] As America's AI-intensive systems may increasingly be composed with allied AI-intensive systems, this becomes a topic for coordination with allies.

(2) Examples of Current Challenges

Unexpected interactions and errors commonly occur in integrated simulations and exercises, illustrating 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.[53]

(3) Recommendations for Adoption

Critical practices to ensure optimal system performance are described in the following non-exhaustive list:

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 blindspots.[54]
      c. Testing for fairness. When testing for fairness, conduct sustained fairness assessments throughout development and deployment and document deliberations made on the appropriate fairness metrics to use. Agencies should conduct outcome and impact analysis to detect when subtle assumptions in the system show up as unexpected and undesired outcomes in the operational environment.[55]
      d. Articulation of performance standards and metrics. Clearly document system performance and communicate to the end user the meaning/significance of such performance metrics.
   2. **Representativeness of the data and model for the specific context at hand.** When using classification and prediction technologies, explicitly consider and document challenges with representativeness of data used in analyses, and the fairness/accuracy of inferences and recommendations made with systems leveraging that data when applied in different populations/contexts.
   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.[56]
   5. **Reliability and robustness:** Employ tools and techniques to carefully bound assumptions of robustness of the AI component in the larger system architecture. Provide sustained attention to characterizing the actual performance envelope (for nominal and off-nominal conditions) throughout development and deployment.[57]

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, which can introduce its own problems.[58] As a priority during testing, challenge (or "stress test") interfaces and usage patterns with boundary conditions and assumptions about the operational environment and use.

B. **Maintenance and deployment**

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

1. **Specifying maintenance requirements** for datasets as well as for systems, given that their performance can degrade over time.[59]
2. **Continuously monitoring AI system performance**, including the use of high-fidelity traces to determine continuously if a system is going outside of acceptable parameters.[60]
3. **Iterative testing and validation**. Training and testing that provide characteristics on capabilities might not transfer or generalize to specific settings of usage; thus, testing and validation may need to be done recurrently, and at strategic intervention points, but especially for new deployments and classes of tasks.[61]
4. **Monitoring and mitigating emergent behavior**. There will be instances where systems are composed in ways not anticipated by the developers, thus, requiring monitoring the actual performance of the composed system and its components.

(4) Recommendations for Future Action

- **Future R&D.** R&D is needed to advance capabilities for TEVV of AI systems to better understand how to conduct TEVV and build checks and balances into an AI system. Improved methods are needed to explore, predict, and control individual AI system behavior so that when AI systems are composed into systems-of-systems their interaction does not lead to unexpected negative outcomes.
- **Metrics**. Progress on a common understanding of TEVV 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.
- **International collaboration and cooperation.** Collaboration is needed to align on how to test and verify AI system reliability and performance, including along shared values (such as fairness and privacy). Such collaboration will be critical amongst allies and partners for interoperability and trust. Additionally, these efforts could potentially include dialogues between the U.S. and strategic competitors on establishing common standards of AI safety and reliability testing to reduce the chances of inadvertent escalation.

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 all will be developed to support humans.

(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 for 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 machine might coordinate.

Designs and methods for human-AI interaction can be employed to enhance human-AI teaming.[62] Methods in support of effective human-AI interaction can help AI systems 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, 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.[63] 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.[64] Other work centers on mechanisms to consider the ideal mix of initiatives, including when and how to rely on human expertise versus on AI inferences.[65] 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 operators.[66] Directions of effort include developing mechanisms for identifying the most relevant information or inferences to provide end users of different skills in different settings.[67] 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[68] shows that confirmation bias can be triggered by the order in which information is 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.

A. **Identification of functions of human in design, engineering, and fielding of AI**
   1. **Define functions, tasks, and responsibilities of human operators and assign them to specific individuals.** Functions will vary for each domain and project, and should be periodically revisited.
   2. **Policies should define the tasks of humans across the AI lifecycle**, given the nature of the mission and current competencies of AI.
   3. **Enable feedback and oversight to ensure that systems operate as they should**.

B. **Explicit support of human-AI interaction and collaboration**
   1. **Human-AI design guidelines**. AI systems designs should take into account 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 accounts for the ways that human and machine behavior can co-evolve;[69] and also avoid automation bias and unjustified reliance on humans in the loop as failsafe mechanisms. Practices should allow for auditing of the human-AI pair. And designs should be transparent to allow for an understanding of how the AI is working day-to-day, supported by an audit trail if things go wrong. Based on context and mission need, designs should ensure usability of AI systems by AI experts, domain experts, and novices, as appropriate.
   2. **Algorithms and functions in support of interpretability and explanation.** Algorithms and functions that provide individuals with task-relevant knowledge and understanding should take into account that key factors in an AI system's inferences and actions can be understood differently by various audiences (e.g., real-time operators, engineers and data scientists, and oversight officials). Interpretability and explainability exists in degrees. 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**. 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 passed from a machine to a human operator and vice versa.

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 and for redress. Infrastructure and instrumentation[70] 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.
6. **Training**. Train and educate individuals responsible for AI development and fielding, including human operators, decision makers, and procurement officers.[71]

(4) Recommendations for Future Action

- **Future R&D.** R&D is needed to advance capabilities 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 in gender and racial minorities. It is also needed to improve human-machine teaming, including disciplines and technologies centered on decision sciences, control theory, psychology, economics (human aspects and incentives), and human factors engineering. 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.
- **Training**. 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. Operators should receive training on the specifics of the system and application, the fundamentals of AI and data science, and refresher trainings (e.g., when systems are deployed in new settings and unfamiliar scenarios, and when predictive models are revised with new data as performance may shift with updates and introduce behaviors unfamiliar to operators).

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 affected by the actions of an AI system should be offered the opportunity to appeal an AI system's determinations. And 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

If a contentious outcome occurs, overseeing entities need the technological capacity to understand what in the AI system caused this. 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 determinations reach the requisite standards to be admitted as evidence. Building the traceability infrastructure to permit auditing (as described in *Engineering Practices*) 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.

1. **Identify responsible actors.** Determine and document the human beings accountable for a specific AI system or any given part of the system and the processes involved. This includes identifying who is responsible for the operation of the system (including its inferences, recommendations, and actions during usage) and who is responsible for enforcing system use policies. Determine and document the mechanism/structure for holding such actors accountable and to whom it should be disclosed for 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 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.[72]
3. **Adopt policies to strengthen accountability.** Identify or, if lacking, establish policies that allow individuals to raise concerns about irresponsible AI development/use, 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 responsible use and fielding requirements; an appealable process for those 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,[73] while also acknowledging the benefits offered by external audits.[74]

4. **External oversight support**. Self-assessment alone may prove to be inadequate in all scenarios. Supporting traceability, specifically documentation to audit trails, will allow for external oversight.[75] Congress can provide a key oversight function throughout the AI lifecycle, asking critical questions of agency leaders and those responsible for AI systems.[76]

(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.

Endnotes

1. 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 visited 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.
2. 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].
3. 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/.
4. 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..
5. International Covenant on Civil and Political Rights, UN General Assembly, United Nations, Treaty Series, vol. 999, at 171 (December 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. *Interim Report*, NSCAI (Nov. 2019), https://www.nscai.gov/reports. Many, but not all nations, share commitments to these values. Even when values are shared, however, they can be culturally relative, for instance, across nations, owing to interpretative nuances.
6. 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 (last accessed June 17, 2020), (https://www.dni.gov/index.php/who-we-are/organizations/clpt/clpt-related-menus/clpt-related-links/ic-principles-of-professional-ethics (on diversity and inclusion).
7. See, e.g., *Privacy Office*, Department of Homeland Security (last accessed June 3, 2020), https://www.dhs.gov/privacy-office#; *CRCL Compliance Branch*, Department of Homeland Security (last accessed May 15, 2020), https://www.dhs.gov/compliance-branch.
8. 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.

9. See *Projects*, U.S. Privacy and Civil Liberties Oversight Board (last accessed June 17, 2020), https://www.pclob.gov/Projects.
10. 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.").
11. 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.
12. 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.").
13. 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.").
14. See, e.g., *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*, Partnership on AI, https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/; 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]; 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; 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.
15. 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. Heidi Ledford, *Millions of Black People Affected by Racial Bias in Health Care Algorithms*, Nature (October 26, 2019), https://www.nature.com/articles/d41586-019-03228-6.

16. See e.g., Dastin, Amazon Scraps Secret AI Recruiting Tool.
17. 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.
18. 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.
19. 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].
20. See Lopez, DoD Adopts 5 Principles.
21. *Model Interpretability in Azure Machine Learning*, Microsoft (July 9, 2020), https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability.
22. Lopez, DoD Adopts 5 Principles.
23. Jessica Cussins Newman, *Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles* (May 5, 2020), Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/ai-decision-points/; 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.
24. Dario Amodei, et al. *Concrete problems in AI safety* (July 2016), https://arxiv.org/abs/1606.06565.
25. Guofu Li, et al., *Security Matters: A Survey on Adversarial Machine Learning*, (Oct. 2018), https://arxiv.org/abs/1810.07339; Elham Tabassi et al., *NISTIR 8269: A Taxonomy and Terminology of Adversarial Machine Learning (Draft)*, National Institute of Standards and Technology (Oct. 2019), https://csrc.nist.gov/publications/detail/nistir/8269/draft.
26. 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.
27. 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"].
28. Id.
29. For 11 categories of attack, and associated overviews, see the Intentionally-Motivated Failures Summary in Kumar, Failure Modes in Machine Learning.

30. 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.
31. 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.
32. Kumar, Failure Modes in Machine Learning.
33. For concerns about generative adversarial networks (GANS) voiced by Gen. Shanahan, JAIC, 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/. Concerns about GANS, information authenticity, and reliable and understandable systems were voiced by Dean Souleles, IC. See *Afternoon Keynote*, 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/.
34. See Lopez, DOD Adopts 5 Principles.
35. 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 an added challenge when different cultural and policy norms are involved when collaborating on development and use with allies.
36. 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.
37. 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.
38. 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.

39. 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/. Documenting caveats of re-use for both datasets and models is critical to avoid "off-label" use harms as one senior official notes. 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/.
40. Jonathan Mace, et al., *Pivot Tracing: Dynamic Causal Monitoring for Distributed Systems*, Communications of the ACM, Vol. 63 No. 3, at 94-102 (March 2020), https://dl.acm.org/doi/10.1145/2815400.2815415 [hereinafter Mace, Pivot Tracing].
41. Aleksander Madry, et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*, MIT (Sept 4, 2019), https://arxiv.org/abs/1706.06083 [hereinafter Madry, Towards Deep Learning Models Resistant to Adversarial Attacks].
42. See e.g., id.; Thomas Dietterich, *Steps Toward Robust Artificial Intelligence*, AI Magazine (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 (June 27, 2016), http://erichorvitz.com/OSTP-CMU_AI_Safety_framing_talk.pdf.
43. On adversarial attacks on ML, see Kevin Eykholt, et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, IEEE Conference on Computer Vision and Pattern Recognition at 1625–1634 (2018), https://ieeexplore.ieee.org/document/8578273; On directions with robustness, see Madry, Towards deep learning models resistant to adversarial attacks. For a more exhaustive list of sources see the Commission's extended version of the Key Considerations in Appendix A-2.
44. Ram Shankar Siva Kumar, et al., *Adversarial Machine Learning--Industry Perspectives*, 2020 IEEE Symposium on Security and Privacy (SP) Deep Learning and Security Workshop (Feb. 2020), https://arxiv.org/pdf/2002.05646.pdf.
45. Dou Goodman, et al., *Advbox: A Toolbox to Generate Adversarial Examples that Fool Neural Networks* (2020), https://arxiv.org/abs/2001.05574.
46. See *What are the Microsoft SDL Practices?*, Microsoft (last accessed July 14, 2020), https://www.microsoft.com/en-us/securityengineering/sdl/practices .
47. 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.
48. Ben Shneiderman, *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*, International Journal of Human–Computer Interaction 2020, Vol. 36, No. 6, at 495–504 (Mar. 23, 2020), https://doi.org/10.1080/10447318.2020.1741118 [hereinafter, Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy].
49. However, test protocols must acknowledge test sets may not be fully representative of real-world usage.
50. 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].

51. One example is "Hidden Feedback Loops", where systems that learn from external world behavior may also shape the behavior they are monitoring. See D. Sculley, et al., *Machine Learning: The High Interest Credit Card of Technical Debt*, Google (2014), https://research.google/pubs/pub43146/.
52. Megha Srivastava, et al., *An Empirical Analysis of Backward Compatibility in Machine Learning Systems*, KDD'20 (forthcoming, Aug. 2020) [hereinafter Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems].
53. David Sculley, et al., *Hidden Technical Debt in Machine Learning Systems*, Proceedings of the 28th International Conference on Neural Information Processing Systems (Dec. 2015), https://dl.acm.org/doi/10.5555/2969442.2969519.
54. 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.
55. 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].
56. Kamar, Combining Human and Machine Intelligence in Large-scale Crowdsourcing.
57. See Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy.
58. Cynthia Dwork, et al., *Individual Fairness in Pipelines*, https://arxiv.org/abs/2004.05167; Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems.
59. *Artificial Intelligence (AI) Playbook for the U.S. Federal Government*, Artificial Intelligence Working Group, ACT-IAC Emerging Technology Community of Interest (Jan. 22, 2020), https://www.actiac.org/act-iac-white-paper-artificial-intelligence-playbook.
60. 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 at 94-102.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.

66. 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.
67. 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.
68. 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.
69. Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy at 495–504.
70. 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.
71. 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.
72. See Raji, Closing the AI Accountability Gap.
73. See id. ("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.
74. 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-new-facial-recognition-algorithm-march/162959/ (highlighting a NIST algorithmic assessment on behalf of U.S. Customs and Border Protection).
75. Raji, Closing the AI Accountability Gap. Maranke Wieringa, *What to Account for When Accounting for Algorithms*, Proceedings of the 2020 ACM FAT Conference, (Jan. 2020), https://doi.org/10.1145/3351095.3372833.

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.

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		