Appendix C: Key Considerations for the Responsible Development and Fielding of Artificial Intelligence (Abridged)

Prefatory Note:

The paradigm and recommended practices described here stem from the Commission's line of effort dedicated to Ethics and Responsible Artificial Intelligence (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 originally in July 2020. Here we present a revised and updated version as part of the Commission's Final Report. Many of the points made here are also reflected in Chapter 7 of the report.

The content herein is an abridged version of the content included in the extended version, which will be featured on NSCAI's website in March 2021 at www.nscai.gov. In the more comprehensive document, we provide additional details and references for technical implementers.

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 has identified five broad categories of challenges and made recommendations for both responsibly developing and fielding AI systems. These recommendations include immediate actions and future work the U.S. government should undertake to help establish best practices to 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 AI systems that are understandable, reliable, robust, and trustworthy.

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 policies and processes to adopt the paradigm's recommended practices, monitor their implementation, and continually refine them as best practices evolve. These recommended practices should apply both to systems that are developed by departments and agencies as well as to 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 in the acquisitions and acceptance processes. As such, the government organization overseeing the bidding process should require that vendors articulate how their practices align with the Key Considerations' recommended practices in their proposals, submissions, and bids.

In each of the five sections 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. America's values, and commitment to upholding them, are reflected in the U.S. Constitution and U.S. laws, regulations, policies, and processes.

One of the seven principles we set forth in the Commission's 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 as well as 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] International treaties that the United States has ratified also demonstrate our values by affirming 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 (IC),[6] the Department of Homeland Security,[7] the Department of Defense (DoD),[8] and oversight entities (e.g., the Privacy and Civil Liberties Oversight Board).[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 (LOAC).[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 concerning autonomy in weapon systems.[13] While not an exhaustive list of U.S. values, the paradigm of considerations and recommended practices for AI that we introduce resonates 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 fielding 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.

(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 for deployment due to concerns with their readiness. Other potential analyses, including factors that lead to success or failure in promotion, can support personnel evaluations. 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 like PNDC 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.

   - An explicit analysis of outcomes that would violate these values should be performed. Policy should prohibit disallowed outcomes that would violate the values above. During system development, analysis of system-specific disallowed outcomes should be performed.[17] As the technology advances, applications evolve, and our understanding of the implications of use grows, these policies should periodically be refreshed.

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.[18] Recommended practices for representing objectives and trade-offs include the following:

1. *Consider and document value considerations in AI systems by specifying how trade-offs with accuracy are handled.* This includes documenting the choices made when selecting operating thresholds that have implications for performance, such as the ratio of true positive and false positive rates or the precision (how many selected items are relevant?) versus recall (how many relevant items are selected?). For example, consider a system designed to recommend if a person entering the U.S. should be pulled aside for more detailed inspection and interview. Precision refers to how many of the people selected for additional processing are valid security concerns; recall refers to how many valid security concerns are flagged for added processing. The trade-off is between allowing a valid security concern to slip past review and detaining persons who are not a security concern. Setting thresholds to increase precision (i.e., reduce the number of persons detained needlessly) will drive down recall (i.e., detain fewer valid security concerns).

2. *Consider and document value considerations in AI systems that rely on representations of objective or utility functions,* especially when assigning weighting that captures the importance of different goals for the system. As an illustration of multiple goals and value weights, consider shopping for a new car. A buyer may identify factors that are important in the decision, such as gas mileage, safety, reliability, and performance. These clearly interact in some cases—for example, gas mileage and performance are likely in tension, and safety is likely correlated partly with vehicle size, which is likely in

tension with gas mileage. When reviewing a set of new cars, the best pick for a buyer will depend on the priorities placed on these factors.

3. *Conduct documentation, reviews, and set limits that reflect disallowed outcomes (through constraints on allowed performance) to ensure compliance with values.*

(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. 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 data sets). 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, for example 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 potential approaches, minimal standards, and engineering proven practices are available.[23]

Additionally, several properties of the engineering 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 frameworks integrate adversarial attacks[25] and unintended faults throughout the lifecycle[26] into a single taxonomy that describes both intentional and unintentional failure modes.[27]

Intentional failures are the result of malicious actors explicitly attacking some aspect of AI system behavior. Taxonomies (e.g., from NIST) on malicious attacks explain the rapidly developing Adversarial Machine Learning (AML) landscape. Attacks span ML training and testing, and each has associated defenses.[28] Categories of intentional failures introduced by adversaries include *training data poisoning* attacks (contaminating training data), *model inversion* (recovering training data used in the model through careful queries), and ML *supply chain attacks* (compromising 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 ML systems.

Examples of unintentional AI failure modes include *reward hacking* (when AI systems learn to achieve a programmed goal in a way that contradicts the programmer's intent) and *distributional shifts* (when a system is tested in one kind of environment but is unable to adapt to changes in other kinds of environments).[30] Another area of failure is the inadequate specification of objectives (as described in Section 1 above on *Representing Objectives and Trade-offs*), leading to unexpected and costly behaviors and outcomes.[31] As AI systems that are separately developed and tested are composed and interact with other AI systems (within one's own services, forces, and agencies, and between U.S. 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 IC must be able to depend on the integrity and security of the data 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 development and fielding of AI systems.

1. *Refine design and development requirements, informed by the concept of operations and risk assessment*, including characterization of failure modes and associated impacts. Conduct systems analysis of operations and identify mission success

metrics and potential functions that can be performed by AI technology. Incorporate early analyses of use cases and scenario development, assess general feasibility and compliance with disallowed outcomes expressed in policy. Critically assess reproducibility (how readily research results can be replicated by a third party) and technical maturity. This includes broad stakeholder engagement and hazard analysis with multidisciplinary experts who ask key questions about potential disparate impacts and document the process undertaken to ensure fairness and the 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. This includes an analysis of the system's potential societal impact and of the impacts of the system's failure modes. 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. *Produce 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 technologies and the 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. *Leverage infrastructure to support traceability, including auditability and forensics.* Invest resources and build capabilities that support the traceability of AI systems. Traceability captures key information about the system's 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. *For security and robustness, address intentional and unintentional failures.*

- *Adversarial attacks and use of robust ML methods.* Expand notions of adversarial attacks to include various ML attacks[41] (as described above) and seek latest technologies that demonstrate the ability to detect and notify operators of attacks and also tolerate attacks (i.e., to enable systems to withstand or to degrade gracefully when targeted by a deliberate attack).[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 system monitoring, failure detection, engineering, and protections during operation. Related efforts and R&D focus on developing and deploying robust AI methods.[43]

- *Adopt a DevSecOps lifecycle for AI systems focused on potential failure modes.* This includes developing and regularly refining threat models to capture and characterize various attacks, establish a matrixed focus for developing and refining threat models, and ensuring DevSecOps addresses ML development, fielding, and when ML systems are under attack.[44]

- *Limit consequences of system failure through system architecture.* Build an overall system architecture that monitors component performance and handles errors when anomalies are detected; build AI components to be self-protecting and self-checking; and include aggressive stress testing under conditions of intended use.

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, assume an offensive posture and use methods to make systems more resistant to adversarial attacks, work with adversarial testing tools, and deploy teams dedicated to trying to break 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 on 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.

- 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.[46] 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 (e.g., weapon system controls).

- *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 tolerate attacks. R&D is also needed to advance capabilities to support risk assessment, including standards, methods, and metrics for evaluating degrees of auditability, traceability, interpretability, explainability, and reliability. For interpretability in particular, R&D is also needed to improve our understanding of the efficacy of interpretability tools and possible interfaces.

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,[47] including assessing its capabilities and blind spots with data representative of real-world scenarios or with simulations of realistic contexts,[48] and its reliability, robustness (i.e., resilience in real-world settings, including withstanding adversarial attacks on AI components), and security during development and deployment.[49] 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 and 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 (e.g., changes in the statistical distribution of data in operation versus model training) 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.[50] A related problem is encountered when the performance of one model in a pipeline changes, degrading the overall pipeline behavior.[51] As America's AI-intensive systems may increasingly be composed and/or interoperable with allied AI-intensive systems, these become important topics 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 necessarily the result of any one component having errors, but rather are due to the interactions of the composed systems.[52]

(3) Recommendations for Adoption

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

A. *Model training and model testing procedures should cover key aspects of performance and appropriate performance metrics.*

1. Use regularly updated standards for testing and reporting of system performance. Standards for metrics and reporting are needed to adequately:
    - a. Achieve consistency across testing and test reporting for critical areas.
    - b. Test for blindspots.[53]
    - c. Test 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.[54]
    - d. Articulate system performance. Clearly document system performance and communicate to the end user the meaning/significance of such performance metrics.

2. *Consider and document the 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. *Evaluate an AI system's performance relative to current benchmarks* where possible. Such benchmarks should assist in determining if a proposed AI system's performance meets or exceeds current best performance.

4. *Evaluate 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.[55]

5. *Provide sustained attention to 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 (for normal and boundary conditions) throughout development and deployment.[56] For systems of particularly high potential consequences of failure, considerable architecture and design work will have been put into making the overall system fail-safe.

6. *For systems of systems, test 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.[57] 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. *Specify maintenance requirements* for datasets as well as for systems, given that their performance can degrade over time.[58]

2. *Continuously monitor and evaluate AI system performanc*e, including the use of high-fidelity traces to determine continuously if a system is going outside of acceptable parameters.[59]

3. *Conduct iterative model 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.[60]

4. *Monitor and mitigate emergent behavior.* There will be instances when 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 persistent and iterative 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 used to assess system performance across attributes for responsible AI according to applications/context profiles. (Such attributes, for example, include fairness, interpretability, reliability, and robustness.) Future work is needed to develop: (1) definitions, taxonomy, and metrics needed to enable agencies to better assess AI performance and vulnerabilities; and (2) metrics and benchmarks to assess reliability and intelligibility of produced model explanations. In the near term, guidance is needed on: (1) standards for testing intentional and unintentional failure modes; (2) exemplar data sets for benchmarking and evaluation, including robustness testing and red teaming; and (3) defining characteristics of AI data quality and training environment fidelity (to support adequate performance and governance).[61]

- *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 among 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 & Teaming

(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. And some systems will serve as more than just support tools and will adopt roles of teammates that actively collaborate with 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 the expected range 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-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 sensitivity to interruption 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 with different skill levels in different settings.[67] Consideration must be given to the prospect of 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 humans in design, engineering, and fielding of AI.

1. Given AI and human capabilities and complementarities, as well as requirements for accountability and human judgment, define the tasks of humans and the goals and mission of the human-machine team across the AI lifecycle. This entails noting needs for feedback loops, including opportunities for oversight.

2. Define functions and responsibilities of humans during system operation and assign them to specific individuals. Functions and responsibilities will vary for each domain and project and should be periodically revisited.

B. Explicit support of human-AI interaction and collaboration.

1. *Extend human-AI design methodologies and guidelines.* AI systems designs should take into account the defined tasks of humans in human-AI collaborations in different scenarios; ensure that 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 (that places unjustified confidence in the results of the computation) and unjustified reliance on humans in the loop as fail-safe 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. *Employ 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. *Design systems to provide cues to human operator(s) about the level of confidence the system has in its results or behaviors.* 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. *Refine 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. *Leverage traceability to assist with system development and understanding.* Traceability processes must capture details about human-AI interaction to retroactively understand where challenges occurred, and why, in order to improve systems and their use 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. *Conduct 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 often performs worse in cases impacting 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 initiative, and passing control between the human and machine. R&D also is needed to optimize the ability of humans and AI to work together to undertake complex, evolving tasks in a variety of environments, as well as for diverse groupings of machines to cooperate with each other, with broader systems, and with human counterparts to achieve shared objectives.

- *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 have the opportunity to appeal an AI system's determinations. Accountability and appellate processes must exist for AI decisions, 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 audit 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. *Appoint full-time responsible AI leads* to join senior leadership. Every department and agency critical to national security and each branch of the armed services, at a minimum, should have a dedicated, full-time responsible AI lead who is part of the senior leadership team. Such leads should oversee the implementation of the Key Considerations recommended practices alongside the department or agency's respective AI principles.

2. *Identify responsible actors.* Determine and document the people 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 development or procurement; operation (including the system's inferences, recommendations, and actions during usage), and maintenance of an AI system, as well as the authorization of a system and enforcement of policies for use. Determine and document the mechanism/structure for holding such actors accountable and to whom it should be disclosed for proper oversight.

3. *Require 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] Policy should establish requirements about information that should be captured about the development process and about system performance and behavior in operation.

4. *Adopt policies to strengthen accountability and governance.* Identify or, if lacking, establish policies that allow individuals to raise concerns about irresponsible AI development/fielding (e.g., via an ombudsman). This requires ensuring a governance structure is in place to address grievances and harms if systems fail, which supports feedback loops and oversight to ensure that systems operate as they should.

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 for 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]

5. *Support external oversight.* Remain responsive and facilitate oversight through documentation processes and other policy decisions.[75] For instance, supporting traceability and specifically documentation to audit trails will allow for external oversight.[76] Self-assessment alone might prove to be inadequate in all scenarios.[77] Congress can provide a key oversight function throughout the AI lifecycle, asking critical questions of agency leaders 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 C - 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*, European Commission at 26-31 (April 8, 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai; *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-assessment*, European Commission (July 17, 2020), https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

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

[3] See Ben Huebner, *Presentation: AI Principles*, Intelligence and National Security Alliance 2020 Spring Symposium: Building an AI-Powered IC (March 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 Disabilities 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 (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. *Interim Report*, NSCAI (Nov. 2019), https://www.nscai.gov/previous-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*, Office of the Director of National Intelligence (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*, Office of the Director of National Intelligence (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*, U.S. Department of Homeland Security (last accessed June 3, 2020), https://www.dhs.gov/privacy-office#; *CRCL Compliance Branch*, U.S. 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 visited June 17, 2020), https://www.pclob.gov/Projects.

[10] See *Department of Defense Law of War Manual*, U.S. Department of Defense (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*, DoD 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 Weapon Systems*, U.S. 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. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazonscraps-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 U.S. 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 (Oct. 26, 2019), https://www.nature.com/articles/d41586-019-03228-6.

[16] See e.g., Dastin, Amazon Scraps Secret AI Recruiting Tool.

[17] This combined approach of stable policy-level disallowed outcomes and system-specific disallowed outcomes is consistent with DoD practices for system safety, for example. See *Department of Defense Standard Practice: System Safety*, U.S. Department of Defense (May 11, 2012), https://www.dau.edu/cop/armyesoh/DAU%20Sponsored%20Documents/MIL-STD-882E.pdf. Depending on the context, mitigating harm per values and disallowed outcomes might entail the use of fail-safe technologies. See Eric Horvitz, *Reflections on Safety and Artificial Intelligence, Exploratory Technical Workshop on Safety and Control for AI* (June 27, 2016), http://erichorvitz.com/OSTP-CMU_AI_Safety_framing_talk.pdf. See also Dorsa Sadigh & Ashish Kapoor, *Safe Control Under Uncertainty with Probabilistic Signal Temporal Logic*, Proceedings of Robotics: Science and Systems XII (2016), https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/RSS2016.pdf.

[18] Mohsen Bayati, et al., *Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study*, PLOS One Medicine (Oct. 8, 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.

Appendix C - Endnotes

[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 (Nov. 16, 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* (April 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 (March 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*, arXiv (July 25, 2016), https://arxiv.org/abs/1606.06565.

[25] Guofu Li, et al., *Security Matters: A Survey on Adversarial Machine Learning*, arXiv (Oct. 23, 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. 11, 2019), https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning [*hereinafter* Kumar, Failure Modes in Machine Learning].

[28] See 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. See also Kumar, Failure Modes in Machine Learning.

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

[30] For more on reward hacking, see Jack Clark, et al., *Faulty Reward Functions in the Wild* (Dec. 21, 2016), https://openai.com/blog/faulty-reward-functions/. For more on distributional shifts, 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 (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.).

[31] Thomas Dietterich & Eric Horvitz, *Rise of Concerns About AI: Reflections and Directions*, Communications of the ACM at 38-40 (Oct. 2015), http://erichorvitz.com/CACM_Oct_2015-VP.pdf. See also Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, Science (Oct. 25, 2019), https://science.sciencemag.org/content/366/6464/447.

[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 (March 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 ML 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 (Aug. 29, 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 data sets 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/allnews/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://m-cacm.acm.org/magazines/2020/3/243034-pivot-tracing/fulltext [*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*, Association for the Advancement of Artificial Intelligence (Fall 2017), https://www.aaai.org/ojs/index.php/aimagazine/article/view/2756/2644; Eric Horvitz, *Reflections on Safety and Artificial Intelligence* (June 27, 2016), http://erichorvitz.com/OSTP-CMU_AI_Safety_framing_talk.pdf.

Appendix C - Endnotes

[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 (June 18-23, 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 *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[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 (May 21, 2020), https://arxiv.org/abs/2002.05646.

[45] Dou Goodman, et al., *Advbox: A Toolbox to Generate Adversarial Examples That Fool Neural Networks* (Aug. 26, 2020), https://arxiv.org/abs/2001.05574.

[46] See *First Quarter Recommendations*, NSCAI (March 2020), https://www.nscai.gov/previous-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.

[47] Ben Shneiderman, *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*, International Journal of Human-Computer Interaction 2020 at 495-504 (March 23, 2020), https://doi.org/10.1080/10447318.2020.1741118 [*hereinafter* Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy].

[48] However, test protocols must acknowledge that test sets may not be fully representative of real-world usage.

[49] 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].

[50] 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/.

[51] Megha Srivastava, et al., *An Empirical Analysis of Backward Compatibility in Machine Learning Systems*, KDD'20 (Aug. 11, 2020), https://arxiv.org/abs/2008.04572 [*hereinafter* Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems].

[52] 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.

[53] Ramya Ramakrishnan, et al., *Blind Spot Detection for Safe Sim-to-Real Transfer*, Journal of Artificial Intelligence Research 67 at 191-234 (Feb. 4, 2020), https://www.jair.org/index.php/jair/article/view/11436.

[54] 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 (April 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].

[55] Kamar, Combining Human and Machine Intelligence in Large-Scale Crowdsourcing.

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

[57] Cynthia Dwork, et al., *Individual Fairness in Pipelines*, arXiv (April 12, 2020), https://arxiv.org/abs/2004.05167; Srivastava, An Empirical Analysis of Backward Compatibility in Machine Learning Systems.

[58] *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.

[59] Ori Cohen, *Monitor! Stop Being A Blind Data-Scientist*, Towards Data Science (Oct. 8, 2019), https://towardsdatascience.com/monitor-stop-being-a-blind-data-scientist-ac915286075f; Mace, Pivot Tracing at 94-102.

[60] 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.

[61] The 2021 NDAA expansion of the National Institute of Standards & Technology (NIST) mission authorizes the standards body to provide such guidance: "National Institute of Standards and Technology Activities (Title LIII, Sec. 5301)—expands NIST mission to include advancing collaborative frameworks, standards, guidelines for AI, supporting the development of a risk-mitigation framework for AI systems, and supporting the development of technical standards and guidelines to promote trustworthy AI systems." Pub. L. 116-283, William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, 134 Stat. 3388 (2021).

[62] Saleema Amershi, et al., *Guidelines for Human-AI Interaction*, CHI '19: Proceedings of the CHI Conference on Human Factors in Computing Systems (May 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*, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Aug. 10-13, 2015), https://www.semanticscholar.org/paper/Intelligible-Models-for-HealthCare%3APredicting-Risk-Caruana-Lou/cb030975a3dbcdf52a01cbd1c140711332313e13.

[64] Eric Horvitz, *Reflections on Challenges and Promises of Mixed-Initiative Interactio*n, AI Magazine (Summer 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 at 52-59 (March 2003), https://cacm.acm.org/magazines/2003/3/6879-models-of-attention-in-computingand-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. An example of co-evolution of machine and human behavior is in ML spam filters. As human spammers determine what characteristics are getting email flagged as spam, they change how they generate spam, which requires the spam-detection models to evolve in a constant "arms race."

[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] Jamie Berryhill, et al., *Hello, World: Artificial Intelligence and Its Use in the Public Sector*, OECD Working Papers on Public Governance (Nov. 21, 2019), https://doi.org/10.1787/726fd39d-en.

Appendix C - Endnotes

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

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

[76] Raji, Closing the AI Accountability Gap.

[77] Brundage, Toward Trustworthy AI Development.

Technical Glossary to the Key Considerations Appendix

This glossary provides a working set of definitions specific to the NSCAI Key Considerations. The Commission acknowledges that the definitions of the terms below may diverge from other scholarly or government definitions and were developed to be accessible to a broad audience.

AI Component: A software object that uses AI, meant to interact with other components, encapsulating certain functionality or a set of functionalities. An AI component has a clearly defined interface and conforms to a prescribed behavior common to all components within an architecture.[1]

AI Lifecycle: The steps for managing the lifespan of an AI system: 1) Specify the system's objective. 2) Build model. 3) Test the AI system. 4) Deploy and maintain the AI system. 5) Engage in a feedback loop with continuous training and updates.[2]

AI System: A system designed or adapted to interact with an anticipated operational environment to achieve one or more intended purposes while complying with applicable constraints and that uses AI to provide a substantial part of its capabilities.[3]

Artificial Intelligence (AI): The ability of a computer system to solve problems and to perform tasks that have traditionally required human intelligence to solve.

Auditability: A characteristic of an AI system in which its software and documentation can be interrogated and yield information at each stage of the AI lifecycle to determine compliance with policy, standards, or regulations.

DevSecOps: Enhanced engineering practices that improve the lead time and frequency of delivery outcomes, promoting a more cohesive collaboration between development, security, and operations teams as they work toward continuous integration and delivery.

Differential Privacy: A criterion for a strong, mathematical definition of privacy in the context of statistical and ML analysis used to enable the collection, analysis, and sharing of a broad range of statistical estimates, such as averages, contingency tables, and synthetic data, based on personal data while protecting the privacy of the individuals in the data.[4]

False Negative: An example in which the predictive model mistakenly classifies an item as in the negative class. For example, a false negative describes the situation in which a junk-email model specifies that a particular email message is not spam (the negative class), when the email message actually is spam, leading to frustration of the junk message appearing in an end user's inbox.[5] In a higher-stakes example, a false negative captures the case in which a medical diagnostic model misses identifying a disease that is present in a patient.

False Positive: An example in which the model mistakenly classifies an item as in the positive class. For example, the model inferred that a particular email message was spam (the positive class), but that email message was actually not spam, leading to delays in an end user reading a potentially important message.[6] In a higher-stakes situation, a false positive describes the situation in which a disease is diagnosed as present when the disease is not present, potentially leading to unnecessary and costly treatments.

High-Fidelity Performance Traces: A commonly used technique useful in debugging and performance analysis. Concretely, trace recording implies detection and storage of relevant events during run-time, for later off-line analysis. High fidelity traces refers to the amount of fine-grained detail captured in the traces.[7]

Human Factors Engineering: The discipline that takes into account human strengths and limitations in the design of interactive systems that involve people, tools and technology, and work environments to ensure safety, effectiveness, and ease of use.[8]

Human in the Loop: The term describes a system architecture in which active human judgment and engagement are part of the operation of a system, and a human is an integral part of the system behavior. An example is the human operator of a remotely piloted vehicle or a decision support system that makes recommendations for a human to decide on.

Human on the Loop: This term describes a system architecture in which a human has a supervisory role in the operation of the system but is not an integral part of the system behavior. An example is an operator monitoring a fleet of warehouse robots—they operate autonomously but can be shut down if the operator determines something is wrong.

Machine Learning (ML): The study or the application of computer algorithms that improve automatically through experience.[9] Machine learning algorithms build a model based on training data in order to perform a specific task, like aiding in prediction or decision-making processes, without necessarily being explicitly programmed to do so.

Model Testing: Testing assesses the performance of a trained model against new, previously unseen inputs, to demonstrate that the model generalizes to produce accurate results beyond just the training data.[10]

Model Training: Training a model simply means learning (determining) good values for all of the internal parameters that determine the model's performance. In supervised learning, for example, a machine learning model is trained by examining many labeled examples and attempting to find a model that minimizes the discrepancies between the real (labelled) values and the values produced by the model.[11]

Technical Glossary to the Key Considerations Appendix

Multi-Party Federated Learning: A machine learning architecture in which many clients (e.g., mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g., a service provider) while keeping the training data decentralized. It can mitigate many of the systemic privacy risks and costs resulting from traditional, centralized machine learning and data science approaches.[12] However, it does introduce new attack vectors that must be addressed.[13]

Precision: A metric for classification models. Precision identifies the frequency with which a model was correct when classifying the positive class. It answers the question "How many selected positive items are true positive?" For example, the percentage of messages flagged as spam that are spam.[14]

Privacy-Preserving AI: Techniques for protecting the privacy of people associated with the training data from adversarial attacks. These techniques include federated learning and differential privacy.[15]

Recall: A metric for classification models. Recall identifies the frequency with which a model correctly classifies the true positive items. It answers the question "How many true positive items were correctly classified?" For example, the percentage of spam messages that were flagged as spam.[16]

Reliable AI: An AI system that performs in its intended manner within the intended domain of use.

Robust AI: An AI system that is resilient in real-world settings, such as an object-recognition application that is robust to significant changes in lighting. The phrase also refers to resilience when it comes to adversarial attacks on AI components.

Run-Time Behavior: The behavior of a program while it is executing (i.e., running on one or more processors).

Trustworthy AI: Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations; (2) it should be ethical, demonstrating respect for, and ensuring adherence to, ethical principles and values; and (3) it should be robust, both from a technical and social perspective, because, even with good intentions, AI systems can cause unintentional harm.[17]

Technical Glossary to the Key Considerations Appendix - Endnotes

[1] See NIST, *NISTIR 7298 Rev. 3, Glossary of Key Information Security Terms* (July 2019), https://csrc.nist.gov/glossary/term/component.

[2] Note that for data-driven AI systems step 2 is expanded and replaced with 2.a) Acquire data to meet the objective, and 2.b) Train the AI system on the data; and these two steps are usually repeated, with data acquisition and training continuing until desired performance objectives are attained. For further discussion on the ML lifecycle, see Saleema Amershi, et al., *Software Engineering for Machine Learning: A Case Study*, IEEE Computer Society (May 2019), https://www.microsoft.com/en-us/research/publication/software-engineering-for-machine-learning-a-case-study/.

[3] See Hilary Sillitto, et al., *Systems Engineering and System Definitions*, International Council on Systems Engineering, (Jan. 8, 2019), https://www.incose.org/docs/default-source/default-document-library/final_-se-definition.pdf.

[4] Kobbi Nissim, et al., *Differential Privacy: A Primer for a Non-technical Audienc*e, Working Group of the Privacy Tools for Sharing Research Data Project, Harvard University, (Feb. 14, 2018), https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf.

[5] See Frank Liang, *Evaluating the Performance of Machine Learning Models*, Towards Data Science (April 18, 2020), https://towardsdatascience.com/classifying-model-outcomes-true-false-positives-negatives-177c1e702810.

[6] *Id.*

[7] See Johan Kraft, et al., *Trace Recording for Embedded Systems: Lessons Learned from Five Industrial Projects*, Runtime Verification at 315-329, https://link.springer.com/chapter/10.1007%2F978-3-642-16612-9_24.

[8] See *Human Factors Engineering*, U.S. Department of Health and Human Services: Agency for Healthcare Research and Quality (Sept. 2019), https://psnet.ahrq.gov/primer/human-factors-engineering.

[9] Thomas M. Mitchell, *Machine Learning*, McGraw-Hill (1997).

[10] See Rob Ashmore, et al., *Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges*, arXiv at 4 (May 2019), https://arxiv.org/abs/1905.04223.

[11] See *Descending into ML: Training and Loss*, Google (last accessed Feb. 15, 2021), https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss.

[12] Peter Kairouz, et al., *Advances and Open Problems in Federated Learning*, arXiv (Dec. 10, 2019), https://arxiv.org/pdf/1912.04977.pdf.

[13] See Vale Tolpegin, et al., *Data Poisoning Attacks Against Federated Learning Systems*, ArXiv (Aug. 11, 2020), https://arxiv.org/abs/2007.08432; Arjun Nitin Bhagoji, et al., *Analyzing Federated Learning Through an Adversarial Lens*, arXiv (Nov. 25, 2019), https://arxiv.org/abs/1811.12470.

[14] See Frank Liang, *Evaluating the Performance of Machine Learning Models*, Towards Data Science (April 18, 2020), https://towardsdatascience.com/classifying-model-outcomes-true-false-positives-negatives-177c1e702810.

[15] For a discussion on how privacy-preserving machine learning works, see Roxanne Heston & Helon Toner, *Have Your Data and Use It Too: A Federal Initiative for Protecting Privacy While Advancing A*I, Day One Project (Jan. 23, 2020), https://www.dayoneproject.org/post/have-your-data-and-use-it-too-a-federal-initiative-for-protecting-privacy-while-advancing-ai; see also Georgios Kaissis, et al., *Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging*, Nature Machine Intelligence at 305-311 (June 8, 2020), https://doi.org/10.1038/s42256-020-0186-1.

[16] See Frank Liang, *Evaluating the Performance of Machine Learning Models*, Towards Data Science (April 18, 2020), https://towardsdatascience.com/classifying-model-outcomes-true-false-positives-negatives-177c1e702810.

[17] See *Ethics Guidelines for Trustworthy AI*, European Commission: High-Level Expert Group on Artificial Intelligence at 5 (April 8, 2019), https://ec.europa.eu/futurium/en/ai-alliance-consultation.