Chapter 7: Establishing Justified Confidence in AI Systems

Blueprint for Action

A Holistic Framework for Ensuring Justified Confidence in AI Systems.

The U.S. Government should align on a common understanding of critical steps needed to ensure justified confidence in AI systems, including confidence in their responsible development and use. The Commission has outlined such a strategy in the *Key Considerations*. The *Key Considerations* provide a framework for the responsible development and fielding of AI that should be adopted by all agencies critical to national security. The framework includes near-term recommendations and topics that agencies should give priority consideration, practices that should be implemented immediately, and policies that should be defined or updated to reflect new AI considerations.

Based on robust feedback from agencies including Department of Defense (DoD), Intelligence Community (IC), Department of Homeland Security (DHS), Federal Bureau of Investigation (FBI), Department of Energy (DoE), Department of State (DoS), and Department of Health and Human Services (HHS), as well as the GSA AI Community of Practice, the *Key Considerations* also outlines areas needing future work and targeted investment to overcome current challenges. Agencies that have already adopted AI principles noted broad alignment between the *Key Considerations* framework and their AI principles. For instance, the framework's recommended practices help operationalize the AI Principles of the DoD and IC[1] and the Principles for Use of AI in Government.[2]

The implementation of the *Key Considerations*' recommendations for future action will be important not only for agencies, but also for cooperation across the world on the responsible development and fielding of AI.[3] Further, while the Commission's mandate led to a focus on recommendations specific to national security entities in our report, many recommendations we elevate in the *Key Considerations* are relevant to the whole country, including other sectors and industry.

Heads of departments and agencies critical to national security should implement the Key Considerations as a framework for the responsible development and fielding of AI systems. Agencies, at a minimum, include the DoD, IC, FBI, DHS, DoE, DoS, and HHS.

Implementing the *Key Considerations* includes developing policies and processes to adopt the framework's recommended practices, monitoring their implementation, and continually refining them as best practices evolve. While this framework covers dozens of practices that contribute toward an ideal state of responsible development and fielding, some practices will be more critical than others depending on the stakes and context, and complying with them will require different costs and resources. This Blueprint for Action provides details on the key actions from this framework that all departments and agencies critical to national security can and should take now as a priority, and investments and resources that the government should make available to further responsible AI across all agencies. *These span recommendations for Robust and Reliable AI; Human-AI Interaction and Teaming; Testing and Evaluation, Verification and Validation; Leadership; and Accountability and Governance.*

Recommendations for Robust and Reliable AI

Recommendation

Action for the Office of Science and Technology Policy (National AI Initiative Office):

- **Focus federal research and development (R&D) investments on advancing AI security and robustness, to help agencies better identify and mitigate evolving AI system vulnerabilities.** Confidence in the robustness and reliability of AI systems requires insight into the development process and the operational performance of the system. Insight into the development process is supported by capturing decisions and development artifacts for review; insight into operational performance is supported by runtime instrumentation and monitoring to capture details of execution. In both development and operation, there is a need to invest in R&D for better tools to facilitate the capture of needed processes and data. R&D should also advance interpretability capabilities to better understand if AI systems are operating as intended. And R&D should support better characterization of performance envelopes to enable the gradual rollout and adoption of AI systems. "Robust AI" is included among the priority research areas found in Chapter 11 of this report.

Action for all Departments and Agencies

- **Create an AI Assurance Framework.** All government agencies will need to develop and apply an adversarial machine learning threat framework to address how key AI systems could be attacked and should be defended. An analytical framework can help to categorize threats to government AI systems, and assist analysts with detecting, responding to, and remediating threats and vulnerabilities.[4] This framework must address supply chain threats to data and models as well as adversarial AI attacks.[5] The framework will support assurance of data authenticity and data and model integrity. "Create an AI Assurance framework" is included among recommendations found in Chapter 1 of this report.

Action for DoD and the Office of the Director of National Intelligence (ODNI):

- **Create dedicated red teams for adversarial testing.** Such red teams should assume an offensive posture, dedicated to trying to break systems and make them violate rules for appropriate behavior.[6] Because of the scarcity of required expertise and experience for AI red teams, the DoD and ODNI should consider establishing enterprise-wide communities of AI red teaming and vulnerability testing capabilities that could be applied to multiple AI developments. The Commission supports the aligned recommendation by WestExec Advisors that the DoD and ODNI should consider "standing up a national AI and ML red team as a central hub to test against adversarial attacks, pulling together DoD operators and analysts, AI researchers, T&E, CIA, DIA, NSA, and other IC components, as appropriate. This would be an independent red-teaming organization that would have both the technical and intelligence expertise to mimic realistic adversary attacks in a simulated operational environment."[7]

Actions for Agencies Critical to National Security:[8]

- **To Meet Baseline Criteria for Robust and Reliable AI —**
    - Upgrade development, procurement, and acquisition strategies to ensure that those accountable for the development, procurement, or acquisition of an AI system (e.g., program managers) adopt the following practices:
        - **Consult an interdisciplinary group of experts to conduct hazard analysis and risk assessments.** These should cover, as relevant to the context: potential disparate impact related to unwanted bias; privacy and civil liberties; international humanitarian law; human rights;[9] system security against targeted attacks;[10] risks of technology being leaked, stolen, or weaponized by adversaries against the U.S.;[11] and steps taken to mitigate identified risks. Agencies should specify in their respective strategies who will consult such a group and who will ultimately make final decisions based on the group's advice.
        - **Improve documentation practices.** Produce documentation describing the data used for training and testing; model(s); other relevant systems (including connections and dependencies within systems); required maintenance (for datasets and models) technical refresh, and when the system is used in a different operational environment. For data, documentation should include how data were sampled, and their provenance. For synthetic data, documentation should also include details on how the data were generated.[12]
        - **Build overall system architectures to limit the consequences of system failure.** Agencies should build an overall system architecture that monitors component performance and handles errors when anomalies are detected; build AI components to be self-protecting (validating input data) and self-checking (validating data passed to the rest of the system); and include aggressive stress testing. As with all high-consequence software systems, where technically feasible, it is important that high-consequence AI systems have overall system architectures that support robust recovery and repair or fail-fast and fail-over to a reliable degraded mode safe system. There should be clear mechanisms for disengaging and deactivating the system when things go wrong.[13]

Recommendations for Human-AI Interaction and Teaming

Recommendation

Action for Department of Defense:

- **Invest in a sustained, multi-disciplinary initiative to enhance human-AI teaming through the Service Laboratories and DARPA.**
    - This initiative should focus on maximizing the benefits of human–AI interaction; better measuring human performance and capabilities when working with AI systems; and helping AI systems better understand contextual nuances of a situation. Advances in human–machine teaming will enable human interactions with AI-enabled systems to move from the current model of interaction where the human is the "operator" of the machine, to a future in which humans are able to have a "teammate" relationship with machines. Specific funding should be dedicated to research on how to improve human–machine teaming and interaction when it involves human life-safety or lethal deployment of a system. Additional research is urgently needed which should address the following issues, among others: delegation of authority, observability, predictability, directability, communication, and trust.
    - R&D investment should also focus on the following:
        - Developing improved human performance assessment, an essential element for AI to understand when and how an appropriate AI intervention should be made.
        - Developing new approaches to humans and AI establishing and maintaining common ground in support of collaboration, particularly cognitive collaboration. This encompasses how a newly established human–AI team scaffolds its mutual understanding and then how it extends it to creatively and collaboratively tackle new challenges.
        - Developing new approaches to trust calibration in human–AI teams. This includes helping people understand when AI is approaching or outside the bounds of its competency envelope, and likewise helping machines understand when people are approaching their limits. The two together will help the human–AI team calibrate trust appropriately and shape their interaction for improved team performance.[14] "Enhanced human–AI interaction and teaming" is included among the priority research areas found in Chapter 11 of this report. This recommendation also maps to the overall DoD R&D funding recommendation in Chapter 3 of this report.

Actions for Agencies Critical to National Security:

- **Meet Baseline Criteria for Effective Human-AI Interaction and Teaming —**
    - National security departments and agencies should clarify policies on human roles and functions, develop designs that optimize human–machine interaction, and provide ongoing and organization-wide AI training.
        - Develop design methodologies that improve our understanding of human–AI interaction and provide specific guidance and requirements that can be assessed.[15] These methodologies should clearly delineate requirements of potential human–AI teaming alternatives and identify whether a proposed solution is likely to meet those requirements or not.

    - Designs should mitigate automation bias (that places unjustified confidence in the results of computation) and unjustified reliance on humans as a failsafe mechanism. They should provide accurate cues to the human operator about the level of confidence the system has in its results/behaviors.

- Ensure policies provide ethical bounds regarding when and where AI is appropriate within a human–AI team in a given context.
    - Policies should identify what functions humans should perform across the AI life cycle; bound assignments and functions, including autonomous functionality; define when tasks should be handed off between a human and machine based on bounds; and require feedback loops to inform oversight and ensure systems operate as expected.
- Provide ongoing training to help the workforce better interact, collaborate with, and be supported by AI systems—including understanding AI tools.[16] As relevant, employees across departments and agencies, and the DoD in particular, should, at a minimum:
    - Gain familiarity with AI tools (e.g., through everyday interaction), including use of AI systems in realistic situations and provide continual feedback to integrate improvements.[17]
    - Receive education that includes fundamentals of AI and data science, including coverage of key descriptors of performance and probabilities.[18]
    - Receive training on interpreting performance standards and metrics correctly and making informed decisions based on them.[19]
    - Gain an understanding of both the fundamental concepts and the high-level concepts in terms of how the system components interact with each other.[20]
    - Have training to recognize human cognitive biases so that human operators interacting with machines can recognize where they might be succumbing to such bias.[21]
    - Receive ongoing refresher trainings suited to system operators. Refresher trainings are appropriate when systems are deployed in new settings and unfamiliar scenarios, and when predictive models are revised with additional training data as system performance may shift, introducing behaviors that are unfamiliar to operators.[22]

Recommendation

Recommendations for Testing and Evaluation, Verification and Validation

Action for the Department of Defense:

- **DoD should tailor and develop TEVV policies and capabilities to meet the changes needed for AI as AI-enabled systems grow in number, scope, and complexity in the Department.[23]**

This should address the following elements:

- **Establish a testing and evaluation, verification and validation (TEVV) framework and culture that integrates testing as a continuous part of requirements specification, development, deployment, training, and maintenance and includes run-time monitoring of operational behavior.[24] An AI testing framework should:**
    - Establish a process for writing testable and verifiable AI requirement specifications that characterize realistic operational performance.[25]
    - Provide testing methodologies and metrics that enable evaluation of these requirements—including principles of ethical and responsible AI, trustworthiness, robustness, and adversarial resilience.[26]
    - Define requirements for performance reevaluation related to new usage scenarios and environments, and distribution over time.
    - Encourage incorporation of operational usage workflow and requirements from the defined use case into the testing.
    - Issue data quality standards to appropriately select the composition of training and testing sets.
    - Support the use of common modular cognitive architectures within suitable application domains that expose standard interface points for test harnessing—supporting scalability through increased automation along with federated development and testing.
    - Support a cyclical DevSecOps-based approach, starting on the inside and working outward, with AI components, system integration, human-machine interfaces, and operations (including human-AI and multi-AI interactions).
    - Remain flexible enough to support diverse missions with changing requirements over time.
- **Extend existing and develop new TEVV methods and tools for dealing with complex, stochastic, and non-stationary systems, including the design of experiments, real-time monitoring of states and behaviors, and the analysis of results.** These methods/tools need to account for human-system interactions (HSI) and their impact on system behavior, system-system interactions and their effect on emergent behavior across a group of systems, and adversarial attacks, via both conventional cyberattacks, and nascent perceptual adversarial AI attacks. Risk assurance concepts should be extended beyond simple "stop-light" charts of consequence and likelihood for a risk being realized and leverage tools that support developing assurance cases that present verifiable claims about system behavior and provide reviewable arguments and evidence to support the claims.[27]
- **Make TEVV tools and capabilities readily available across the DoD**, including downloadable and configurable AI TEVV software stacks.[28] In addition, the DoD should ensure tools that support TEVV and reliability and robustness goals are available department-wide including tools for bias detection, explainability, and documentation across the product life cycle (e.g., of data inputs and system outputs).
- **Update existing and create new live, virtual, and constructive test ranges for AI-enabled systems (blending modeling and simulation, augmented reality, and cyber physical system environments).** Upgraded test ranges should include live-virtual-constructive environments, the ability to capture data from testing, and the ability to evaluate data from operations. They should support: 1) The full exploration of potential system states and behaviors over a range of runtimes and fidelity levels;

2) the co-development of AI-system functionality and concepts of operations (CONOPS) associated with human-system and system-system teaming; and 3) a fuller understanding of the impact of adversarial activities undertaken to counter these systems. Build these capabilities upon extensive modeling and simulation (M&S) facilities, human and constructive adversarial "red teams," virtual and augmented reality enablers, full instrumentation, and post-run big data analytics capability.

- **Support the T&E community by restructuring the processes that underlie requirements specification, system design, T&E itself, and CONOPS development.** This includes continuing DoD investments and policies supporting architecting software-intensive systems using common frameworks and composable subsystems,[29] the inclusion of runtime instrumentation (adding the capture of internal states of the system, analogous to a flight data recorder on aircraft) in system design and monitoring during operation,[30] the proper curation and protection of data used in training these systems, and a heavy investment in successively sophisticated M&S, starting at the requirements stage and proceeding through development, TEVV, and operator training.

Action for the National Institutes of Standards and Technology (NIST):

- **NIST should provide and regularly refresh a set of standards, performance metrics, and tools for qualified confidence in AI models, data, and training environments, and predicted outcomes.**[31] Over time, as the science of how to test systems across responsible AI attributes evolves, NIST should provide guidance on:
  - Metrics to assess system performance per responsible AI attributes (e.g., fairness, interpretability, reliability, robustness) and according to application/context profiles. This should include:
    - Definitions, taxonomy, and metrics needed to enable agencies to better assess AI performance and vulnerabilities.
    - Metrics and benchmarks to assess reliability of model explanations.[32]
  - For each of the metrics and technical measures created, NIST should also provide measurable outcomes against which success can be determined.[33]
- **In the near term, NIST should also provide guidance on:**
  - Standards for testing intentional and unintentional failure modes
  - Exemplar data sets for benchmarking and evaluation, including robustness testing and red teaming
  - Defining characteristics of AI data quality and training environment fidelity (to support adequate performance and governance)

In conducting the above, NIST should publish quarterly updates to inform departments and agencies about the trustworthy frameworks, standards, and metrics work it is planning.[34]

Action for the Office of Science and Technology Policy - National AI Initiative Office:

- **The federal government should increase R&D investment to improve our understanding of how to conduct TEVV.** This is needed to better understand how to efficiently and effectively test AI systems to provide objective assurance to support a justified level of confidence, build checks and balances in systems, and how to monitor and mitigate unexpected behavior in a composed system-of-systems or when systems interact. Such R&D should advance our understanding of how to test system performance across responsible AI attributes (e.g., fairness, interpretability, reliability, and robustness). This recommendation is echoed by the priority research areas found in Chapter 11 of this report, including "TEVV of AI Systems" and "standard methods and metrics for evaluating degrees of auditability, traceability, interpretability, explainability, and reliability." For more information, see also Chapter 3 of this report.

Actions for Agencies Critical to National Security:

- **To ensure optimal performance of AI systems, national security departments and agencies should:**
    - Plan for and execute aggressive stress testing of AI components to evaluate error handling and robustness against unintentional and intentional threats under conditions of intended use.
    - Include testing for blind spots and fairness throughout development and deployment. Testing and validation should be done iteratively at strategic intervention points, especially for new deployments.
    - Clearly document system performance requirements (including identified system hazards), metrics used for TEVV, deliberations on the appropriate fairness metrics to use, and the representativeness of the test data for the anticipated operational environment.
    - Conduct red teaming to rigorously challenge AI systems, exploring their risks, limitations, and vulnerabilities including intentional and unintentional failure modes.

Recommendations for Leadership

Recommendation

Actions for DoD, IC, FBI, DHS, DoE, DoS, and HHS:

- **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. Responsible AI Leads must have dedicated staff, resources, and authority to succeed in their roles. Every lead should have at least two full-time staff to effectively fulfill the following:**
    - The Responsible AI Lead in each department should oversee the implementation of the *Key Considerations* recommended practices alongside the department/agency's respective AI principles.[35] This includes driving policy development and training programs for the department and internally coordinating Responsible AI leads in the department's supporting branches or agencies (as applicable) to ensure synergistic implementation of such policies and programs. The department lead should determine the Responsible AI governance structure to ensure centralized and consistent policies[36] are applied across the department.

- The department Responsible AI Lead and those supporting Responsible AI leads should collectively:
  - provide Responsible AI training to relevant personnel;
  - serve as subject matter experts regarding existing and proposed Responsible AI policy and best practices;
  - shape procurement policy and guidance for product managers to ensure alignment with recommended practices and adopted AI principles;
  - build a central repository of Responsible AI work going on in the department, and lessons learned from practical implementation across the department, to help streamline department efforts;
  - ensure interagency knowledge sharing for responsible AI, including iterative sharing of best practices, resources and tools, evolving risks and vulnerabilities, and other lessons learned from practical implementation;
  - annually produce a report for Congress on department resources received, any additional resources needed, and an update on required policy work and implementation of recommended practices.
- Where possible, centralized assessments and shared learnings should be communicated across a department's elements or branches, to avoid units spending unnecessary and duplicative resources and to accelerate practices that reduce friction in workflows. Responsible AI Leads in each department should consider the Learning, Knowledge, and Information Exchange (LKIE) framework as a way to accelerate organizational knowledge within their department given the need to leverage collective insights that are gleaned from on-the-ground experience where the *Key Considerations* will be put into practice rather than letting the insights sit in silos.[37] Furthermore, having Responsible AI "champions"[38] who "socialize" this knowledge can help to transfer the knowledge within and across different U.S. Government agencies and components.[39]
- Borrowing from the world of cybersecurity, the Lead also should consider coordinating the adoption of an empirically driven prioritization matrix for risk management.[40]

Action for the National AI Initiative Office:

- **In addition to the National AI Initiative responsibilities defined in the National Defense Authorization Act for Fiscal Year 2021 (FY2021 NDAA),[41] the Office should create a standing body of multi-disciplinary experts who can be voluntarily called upon by agencies as a resource to provide advice on Responsible AI issues.** The group should include people with expertise at the intersection of AI and other fields such as ethics, law, policy, economics, cognitive science, and technology including adversarial AI techniques. As the government upskills and diversifies its workforce with AI expertise, this standing body of experts should help fill gaps in multi-disciplinary expertise that can be called upon by agencies as needed for processes including multi-disciplinary risk assessment, human-AI teaming assessments, and red-teaming.
- Leveraging this in-house expertise, and serving as the central resource for best practice sharing across agencies, it should also:

- Maintain a Learning, Knowledge, and Information Exchange (LKIE) repository to benefit all agencies:
  - A repository compiling insights across agencies (e.g., per the LKIE framework mentioned above) would accelerate organizational knowledge and support the goal of interagency sharing of insights gleaned from on-the-ground practice—rather than letting such insights sit in silos.[42] These collective insights would be generalized from bright spots of successful AI adoption and from lessons learned from AI adoptions that faced problems in development or use.[43] Centralized insights will also provide a resource to help agencies address critical questions that will arise as AI capabilities evolve. Examples of potential critical questions include how to support redress with updated policies and procedures; how to efficiently monitor behavior in operation; and how to effectively measure and address changes introduced by technical refresh. With technical refresh, it is necessary to analyze results carefully. Even if overall performance may be steady or improve after a refresh, the aggregate performance can mask certain parts of the performance envelope where results are significantly skewed and problematic.

Action for Congress:

- **To enable departments and agencies critical to national security to execute Responsible AI work department-wide, and to encourage necessary appointments of Responsible AI personnel, Congress should appropriate an estimated $21.5 million each fiscal year to fund billets.**
  - Organizations that have high mission complexity and diverse components may need more support staff and/or Responsible AI Leads to be allocated across the organization. The Commission recommends that, at a minimum, the following is needed:
    - For the DoD, a department-wide Responsible AI (RAI) Lead and supporting RAI Leads for each branch of the armed services, with each lead supported by two staff members;
    - For the Intelligence Community, an ODNI RAI Lead and supporting RAI Leads for each IC agency, with each lead supported by two staff members;
    - For the DOE, a RAI Lead and a supporting RAI Lead for the National Laboratories, with each lead supported by two staff members; and
    - For the FBI, DHS, and HHS, a RAI Lead in each respective organization who is supported by two staff members.[44]

Recommendations for Accountability and Governance

Recommendation

Actions for Agencies Critical to National Security:

- **Adapt and extend existing policies to ensure accountability is established and documented across the AI life cycle for any given AI system and its components.[45]**
- **Establish clear requirements about information that should be captured about the development process[46] (via traceability) and about system performance and**

behavior in operation (via runtime monitoring) to support reliability and robustness as well as auditing for oversight. Instrumentation to support monitoring can contribute to insights about system performance, but must be provided thoughtfully to prevent new openings for external espionage or tampering with AI systems.[47]

- Guidance should include technical audit trail requirements per mission needs for high-stakes systems.

- **Institute comprehensive oversight and enforcement practices.**
  - Agencies should identify or establish new policies, due to the novelty and advancement of AI technologies, that:
    - allow individuals to raise concerns about irresponsible AI development (e.g., through an ombudsman); and
    - provide layers of human oversight or redundancy so that high-stakes decisions do not rely entirely on determinations made by the AI system.[48]
  - Adapt and extend oversight practices to include reporting requirements[49] for AI systems; a mechanism to allow for thorough review of the most sensitive and high-risk AI systems (to ensure auditability and compliance with deployment 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.[50]
  - Establish selection criteria that indicate if and when specific recommended practices (as found in the *Key Considerations*) need to be used according to system and mission risks.
  - Define triggers that would require escalated review of an AI system.

Blueprint for Action: Chapter 7 - Endnotes

[1] See *Key Considerations for Responsible Development & Fielding of Artificial Intelligence Supporting Visuals*, NSCAI (July 2020), https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-Supporting-Visuals.pdf.

[2] See Donald J. Trump, *Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*, The White House (Dec. 3, 2020), https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/. The Principles for Use of AI in Government do not apply to national security agencies; however, they do apply to agencies the Commission considers critical for national security (e.g., Department of State and Department of Health and Human Services).

[3] See the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI*. For additional details on international cooperation, see the Commission's recommendation for future action in the sections on "Aligning Systems and Uses with American Values and the Rule of Law" and "System Performance" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[4] There are various public and private efforts ongoing. See for instance the MITRE-Microsoft adversarial ML framework, Ram Shankar Siva Kumar & Ann Johnson, *Cyberattacks Against Machine Learning Systems Are More Common than You Think*, Microsoft Security (Oct. 22, 2020), https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think/; *Adversarial AI Threat Matrix: Case Studies*, MITRE (last accessed Jan. 10, 2021), https://github.com/mitre/advmlthreatmatrix/blob/master/pages/case-studies-page.md.

[5] *NISTIR 8269 (Draft): A Taxonomy and Terminology of Adversarial Machine Learning*, National Institute of Standards of Technology (October 2019), https://csrc.nist.gov/publications/detail/nistir/8269/draft.

[6] See the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI*. For additional details on the Commission's recommendation for red teaming, see the section on "Engineering Practices" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[7] See Michele Flournoy, et al., *Building Trust Through Testing* (October 2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

[8] As noted above, the Commission considers these, at a minimum, to include the DoD, IC, DHS, FBI, DoE, Department of State, and HHS.

[9] 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 (October 2018), https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

[10] These can include reidentification attacks. Departments and agencies should use privacy protections such as robust anonymization that can withstand sophisticated reidentification attacks, and when possible, privacy-preserving technology such as differential privacy, federated learning, and ML with encryption of data and models.

[11] 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.

[12] Such documentation should support assurances of the authenticity, integrity, and provenance of data.

[13] See *Making Responsible AI the Norm Rather than the Exception*, Montreal AI Ethics Institute at 9 (Jan. 13, 2021), https://arxiv.org/pdf/2101.11832.pdf [*hereinafter* MAIEI Report] (This includes "building fail safes and backup modes that don't have to rely on continuous access to the 'intelligent' elements and have graceful failures that minimize harm.").

[14] See Brian Wilder, et al., *Learning to Complement Humans*, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (2020), https://www.ijcai.org/Proceedings/2020/0212.pdf.

[15] For an example of applicable guidelines, see 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.

[16] For more on training, see the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI.* For additional details on the Commission's recommendation for training, see the section on "Human-AI Interaction and Teaming" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[17] Such everyday interaction and continual feedback loops will further enhance TEVV.

[18] See the Appendix of this report containing the abridged version of *NSCAI's Key Considerations for Responsible Development & Fielding of AI.* For additional details on the Commission's recommendation for training, see the section on "Human-AI Interaction and Teaming" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission). See also MAIEI Report at 7.

[19] MAIEI Report at 7.

[20] Id.

[21] Id.

[22] See the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI.* For additional details on the Commission's recommendation for training, see the section on "Human-AI Interaction and Teaming" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[23] To the greatest extent possible, DoD should develop TEVV policies and capabilities in coordination with the Office of the Director of National Security.

[24] To achieve this, heavy investment is needed that supports requirements generation/traceability, the integration of heterogeneous test data at all stages of testing, and the use of extensive M&S, test automation, and data analytics wherever feasible.

[25] This should be framed broadly, providing left/right limits that provide guidance but do not limit innovation.

[26] These testing methodologies and metrics should support robust red teaming, meeting the DoD's particular needs for solutions hardened to adversarial actions.

[27] Miles Brundage, et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, arXiv (April 20, 2020), https://arxiv.org/abs/2004.07213.

[28] TEVV tools and software stacks should be shared across the Department using the AI Digital Ecosystem described in Chapter 2 of this report.

[29] Such frameworks for composing testable AI systems should be established and accessed through the AI Digital Ecosystem described in Chapter 2 of this report.

Blueprint for Action: Chapter 7 - Endnotes

[30] See e.g., Software Acquisition Pathway Interim Policy and Procedures, Memorandum from the Under Secretary of Defense, to Joint Chiefs of Staff and Department of Defense Staff (Jan. 3, 2020), https://www.acq.osd.mil/ae/assets/docs/USA002825-19%20Signed%20Memo%20(Software).pdf (stating that program managers are required to "achieve ... continuous runtime monitoring of operational software").

[31] This recommendation is in line with Congress' expansion of NIST's mission regarding AI standards in the FY2021 NDAA, section 5301 to include: "advance collaborative frameworks, standards, guidelines" for AI, "support the development of a risk-mitigation framework" for AI systems, and "support 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).

[32] "Documentation of the assumptions and limitations of the benchmarks so created will also be essential in helping those utilizing them to make sure they will get the intended intelligence from it rather than becoming falsely confident about the system." MAIEI Report at 9.

[33] MAIEI Report at 9.

[34] Doing so will enable departments and agencies to plan and prioritize any internal standards work accordingly (e.g., avoiding redundant or obsolete efforts).

[35] For each of the metrics and technical measures mentioned in the Key Considerations, it will be important to have measurable outcomes against which success can be determined. See MAIEI Report at 9.

[36] This includes, for example, "Accountability and Governance" policy work identified below in this Blueprint for Action.

[37] MAIEI Report at 11-16.

[38] "AI champions" are a cross-functional group of ambassadors, who can, for example, consider ways to operationalize AI ethical principles and serve as internal advocates and evangelists for responsible AI. See *Department of Defense Joint Artificial Intelligence Center Responsible AI Champions Pilot*, DoD (last accessed Feb. 3, 2021), https://www.ai.mil/docs/08_21_20_responsible_ai_champions_pilot.pdf; Tim O'Brien, et al., *How Global Tech Companies can Champion Ethical AI*, World Economic Forum (Jan. 14, 2020), https://www.weforum.org/agenda/2020/01/tech-companies-ethics-responsible-ai-microsoft/.

[39] MAIEI Report at 12.

[40] MAIEI Report at 20-23.

[41] Pub. L. 116-283, Div. E., Title LI, sec. 5102, William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, 134 Stat. 3388 (2021).

[42] MAIEI Report at 11-16.

[43] For instance, this could include communication of failure modes (e.g., when a system produces a formally correct, but unsafe outcome), and instances to establish a shared understanding of how and where the systems go wrong. Leveraging this, agencies should tap into USG network-wide expertise to address those failures. See Ram Shankar Siva Kumar, et al., *Failure Modes in Machine Learning*, Microsoft (Nov. 11, 2019), https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning; MAIEI Report at 7.

[44] Collectively, considering both Responsible AI Leads and supporting staff, this recommendation proposes 21 full-time employees (FTEs) for the DoD; 54 for the IC; 3 for the FBI; 3 for DHS; 6 for DoE; 3 for HHS; and 3 for DoS.

[45] As noted in the *Key Considerations*, agencies should determine and document who is accountable for a specific AI system or any given part of an AI system and the processes involved with it. This should identify 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. See the Appendix of this report containing the abridged version of NSCAI's Key Considerations for Responsible Development & Fielding of AI. For additional details on the Commission's recommendation for accountability, see the section on "Accountability and Governance" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[46] For a list of recommended information that documentation should note about system development, see the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI*. For additional details on the Commission's recommendations for traceability, see the *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[47] For example, "APIs are 'doors' to access digital infrastructures thus, the security and resilience of digital environments will also depend on the robustness of the API infrastructure." V. Lorenzino, et al., *Application Programming Interfaces in Governments: Why, What and How*, European Union Joint Research Centre (2020). https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/application-programming-interfaces-governments-why-what-and-how.

[48] See Frances Duffy, *Ethical Considerations for Use of Commercial AI*, John Hopkins Applied Physics Laboratory at 31 (Dec. 2020). For example, DoD Directive 3000.09 requires human oversight in the targeting and execution process for lethal autonomous weapons. See *DoD Directive 3000.09: Autonomy in Weapons Systems*, U.S. Department of Defense (May 8, 2017), https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[49] For example, reporting risk and impact assessment, steps taken to mitigate such risks, and system performance during testing and fielding.

[50] As with all consequential software systems, developers and adopters of consequential AI systems must adapt and extend existing support for oversight, audit, reporting, and appealable accountability for developing or using systems irresponsibly, and a redress process where appropriate for those affected by system actions. Existing frameworks must be tailored to reflect issues of concern with AI-based systems (particularly based on machine learning). These issues of concern are discussed in more detail in the Appendix of this report containing the abridged version of NSCAI's *Key Considerations for Responsible Development & Fielding of AI*. For additional details on the Commission's recommendations for accountability and governance, see the *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission)