DEFENSE
INNOVATION UNIT

RESPONSIBLE AI GUIDELINES IN PRACTICE

LESSONS LEARNED FROM THE DIU'S PORTFOLIO

BY: JARED DUNNMON, BRYCE GOODMAN,
PETER KIRECHU, CAROL SMITH, & ALEXANDREA VAN DEUSEN

ACCELERATING COMMERCIAL TECHNOLOGY FOR NATIONAL SECURITY

ABOUT THE AUTHORS

Jared Dunnmon is the Technical Director for Artificial Intelligence and Machine Learning at the Defense Innovation Unit.

Bryce Goodman is the Chief Strategist for Artificial Intelligence and Machine Learning at the Defense Innovation Unit.

Peter Kirechu is the Portfolio Analyst for Artificial Intelligence and Machine Learning at the Defense Innovation Unit.

Carol Smith is a Senior Research Scientist in human-machine interaction at Carnegie Mellon University's Software Engineering Institute.

Alexandrea Van Deusen is a Design Researcher at Carnegie Mellon University's Software Engineering Institute.

ACKNOWLEDGEMENTS

The Authors thank Ashley Moore, U.S. Agency for Global Media; Brandie Nonnecke, University of California-Berkeley; Brian Drake, Defense Intelligence Agency; David Bray, Atlantic Council; David Kuehn, Department of Transportation; Helen Toner, Center for Security and Emerging Technology; Jessica Young, Office of Science and Technology Policy; John Stockton, Quantifind; Leo Meister, Department of Health and Human Services; Lynne Parker, Office of Science and Technology Policy; Michael Kanaan, USAF Accelerator; Rama Elluru, National Security Commision on Artificial Intelligence; and Shankar Rao, Enlitic for providing valuable feedback on earlier drafts of this report. Their assistance does not imply any endorsement of this report's contents or recommendations.

DIU collaborated with members of the Software Engineering Institute (SEI) on this work. The SEI serves the nation as a not-for-profit, Federally Funded Research and Development Center (FFRDC), specifically established by the DoD to focus on software and cybersecurity. The SEI currently leads a community-wide movement to mature the discipline of AI Engineering for Defense and National Security. Based at Carnegie Mellon University, the SEI is a global research university annually rated among the best for computer science and engineering programs.

ABOUT THE DEFENSE INNOVATION UNIT

The Defense Innovation Unit (DIU) strengthens our national security by accelerating the adoption of commercial technology throughout the military. DIU partners with organizations across the Department of Defense (DoD) and with "non-traditional" companies to field advanced commercial solutions that address national security challenges. This mission demands a deep understanding of commercial technology and the driving factors that influence current and future market trends. DIU is organized around six portfolios that focus on mission areas where commercial technologies can best advance DoD technology needs. These focus areas include Advanced Energy and Materials, Artificial Intelligence, Autonomy, Cyber, Human Systems, and Space. The DIU workforce is composed of subject matter experts with multidisciplinary experience in the commercial, defense, technology, and U.S. Government policy sectors.

DIU leverages Other Transaction (OT) authority to direct-award prototype agreements. Successful prototypes may result in the direct award of a follow-on Production agreement without the use of further competitive procedures.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

As part of its mission to accelerate adoption of commercial technology within the Department of Defense (DoD), the Defense Innovation Unit (DIU) launched a strategic initiative in March 2020 to integrate the DoD's Ethical Principles for Artificial Intelligence (AI) into its commercial prototyping and acquisition programs. Drawing upon best practices from government, non-profit, academic, and industry partners, DIU explored methods for implementing these principles in several of its AI prototype projects. The result is a set of Responsible Artificial Intelligence (RAI) Guidelines.

The RAI Guidelines consist of specific questions that should be addressed at each phase in the AI lifecycle: planning, development, and deployment. They provide step-by-step guidance for AI companies, DoD stakeholders, and program managers to ensure AI programs align with the DoD's Ethical Principles for AI and ensure that fairness, accountability, and transparency are considered at each step in the development cycle. DIU is actively deploying the RAI Guidelines on a range of projects that cover applications including predictive health, underwater autonomy, predictive maintenance, and supply chain analysis.

It is important to note that the RAI Guidelines cannot offer universally reliable ways to "fix" shortcomings such as biased data, inappropriately selected algorithms, or poorly defined applications in every situation. Furthermore, some systems that are proposed for national security use cases may have no route to responsible deployment—deciding not to pursue an AI capability should be an acceptable outcome of adhering to the RAI Guidelines. Finally, the RAI Guidelines should be viewed as complementary to the internal ethics review and related testing and evaluation (T&E) procedures that many companies providing AI products have in place.

Over the course of applying these RAI Guidelines to active programs and iterating on their content, we have identified several key learnings for each phase of the AI development lifecycle:

- **Planning:** define the task, success metric, and baselines appropriately; obtain access to data required to support the capability; identify stakeholders and mission owners; conduct detailed harms modeling; prescribe processes to safely address system errors and revert malfunctioning systems back to a previously functioning version.
- **Development:** ensure that developers take steps to mitigate the potential negative impact of data or model manipulation; delineate metrics and indicators included in post-deployment monitoring; assign the authority to make changes to the capability to a specific, accountable party; design the system interface to give users the ability to understand how outputs are produced; and establish plans for routine system auditing.
- **Deployment:** conduct continuous task and data validation to ensure task specification and data inputs remain valid and secure; confirm that new data does not degrade system performance; leverage functional testing to evaluate whether the capability still performs the desired task sufficiently well to be operationally useful; and include harms assessment and quality control steps to make certain that potential negative impacts on stakeholders are constantly reassessed and mitigated where necessary.

The RAI Guidelines are a useful starting point for operationalizing the DoD's Ethical Principles for AI. DIU will continue collaborating with experts and stakeholders from government, industry, academia, and civil society to further develop the RAI Guidelines.

INTRODUCTION

In March 2020, the Defense Innovation Unit (DIU) launched a strategic initiative to operationalize the Ethical Principles for Artificial Intelligence (AI) officially adopted by the Department of Defense (DoD) on February 24, 2020.[1] While the Principles do not prescribe a methodology or offer concrete directions, they identify a clear need for practical implementation guidelines for the technology development and acquisition workforce.

For more than a year, DIU explored methods for implementing these principles with DoD partners in several AI prototype projects that cover applications including, but not limited to, predictive health, underwater autonomy, predictive maintenance, and supply chain analysis. The result is a set of Responsible AI Guidelines (hereafter referred to as "RAI Guidelines") that are informed by DIU's practical experience, but also draw upon best practices from government, non-profit, academic, and industry partners.

DoD Ethical Principles for Artificial Intelligence[2]

Responsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

Equitable. The Department will take deliberate steps to minimize unintended bias in AI capabilities.

Traceable. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

Reliable. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

Governable. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

> **"DIU's RAI Guidelines provide step-by-step guidance for AI companies, DoD stakeholders, and program managers on how to ensure AI programs are built with principles of fairness, accountability and transparency at each step in the development cycle."**

The RAI Guidelines were inspired by the requirements set forth by the Deputy Secretary of Defense in a May 27, 2021 memorandum directing DoD officials to "develop tools, policies, processes, systems, and guidance" that ensure that AI technology systems comply with ethical development principles as part of the Department's acquisition policies.[3] DIU's RAI Guidelines provide step-by-step guidance for AI companies, DoD stakeholders, and program managers to ensure AI programs reflect the DoD's Ethical Principles for AI and that fairness, accountability, and transparency are considered at each step in the development cycle. In addition, the Guidelines will support the Department's ability to meet 2021 Responsible AI mandates put forth by both Congress and Deputy Secretary of Defense Kathleen Hicks in a timely manner. Specifically, the Guidelines address a key requirement in the FY21 National Defense Authorization Act (NDAA) which directed the Secretary of Defense to ensure that the DoD has "the ability, requisite resourcing, and sufficient expertise to ensure that any artificial intelligence technology...is ethically and reasonably developed."[4]

This paper provides a summary of the RAI Guidelines that have resulted from DIU's efforts to operationalize the DoD's Ethical Principles for AI within its prototyping efforts. It also provides detailed case studies demonstrating the value of the RAI Guidelines in practice while identifying specific lessons learned from these efforts. Lastly, this paper is accompanied by a set of instructive materials in the Appendix that can be used by personnel across DoD to apply DIU's RAI Guidelines to their own technology development or acquisition programs, or to inform future research in ethical AI.

[1] "Ethical Principles for Artificial Intelligence." United States Department of Defense. https://www.ai.mil/docs/Ethical_Principles_for_Artificial_Intelligence.pdf

[2] Ibid.; Note that DIU uses "planning, development, and deployment" as the three phases of the process by which AI systems are built in this document as opposed to the "development, deployment, and use" language used in the DoD's Ethical Principles for AI. We found "planning, development, and deployment" to more directly align with the commercial software engineering process and DIU's commercial solutions opening process, so have chosen to use that framing in this document.

[3] Deputy Secretary of Defense, "Implementing Responsible Artificial Intelligence in the Department of Defense." 1-3. https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

[4] H.R.6395 - William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, (01/01/2021). 116th Congress (2019-2020). https://www.congress.gov/bill/116th-congress/house-bill/6395

BACKGROUND: DIU'S RESPONSIBLE AI GOALS

With increased attention on the ethical aspects of developing and deploying AI systems, it became incumbent upon DIU to provide guidance on how the DoD's Ethical Principles for AI could effectively and efficiently be put into practice in new and ongoing DIU projects. Moreover, actionable guidance needed to be approachable and relevant not only to the companies providing AI solutions, but also to DIU's DoD partners (which include headquarters and acquisition organizations as well as end users) and the DIU program managers who facilitate prototype projects.

DIU sits in a unique position as the only organization within the DoD exclusively focused on accelerating the adoption of existing commercial technology across the U.S. military. Part of DIU's mission is to ensure that the most advanced commercial technologies get into the hands of DoD personnel as quickly, efficiently, and responsibly as possible. The majority of AI companies working with DIU are new to doing business with the Federal government and, in large part, are open to taking the DoD on as a customer because of the more agile and "commercial friendly" acquisition processes DIU offers.[5] Therefore, DIU was invested in discovering ways to operationalize AI ethically without compromising this value proposition to commercial companies or our ability to deliver leading-edge technology to the Department at the speed of relevance.

> **"DIU approached the development of RAI Guidelines with a set of tenets in mind to ensure our continued ability to maximize benefits for national security while aligning with American laws, norms, and values."**

To date, many of the most egregious examples of unethical AI—such as bias in facial recognition systems or pedestrian deaths from autonomous vehicles—are not always the result of failing to follow a particular policy or ethical code; in reality, they are often caused by insufficiently precise problem formulation or poor engineering, program management, and monitoring practices.[6] Given this context, DIU approached the development of RAI Guidelines with a set of tenets in mind to ensure our continued ability to maximize benefits for national security while aligning with American laws, norms, and values.

RAI FOUNDATIONAL TENETS

- **Actionable:** The RAI Guidelines are not an abstract policy instrument, but a tool for applying best practices drawn from DIU's experience building and scaling AI across the DoD, as well as input from AI practitioners in academia and industry. The intended audience is project execution teams, including vendors, users, acquisition officers, and program offices that are ultimately responsible for project success.
- **Concrete:** If the intent is to change the way that AI technologies are evaluated, selected, prototyped, and adopted, a list of questions or high-level principles is insufficient. Thus, every step in the RAI Guidelines is accompanied by instructive commentary, guidance, and resources that explain why specific questions are asked and what concerns the responses should address.
- **Realistic:** The aim is not to guarantee the best possible outcome, but to avoid potentially bad outcomes by leveraging the best tools at our disposal. As such, the Guidelines should be updated over time to reflect the state of the art in technology, best practices, and contemporary ethics.
- **Adaptive:** The RAI Guidelines are intended to be applicable to any project that involves AI, whether that be for predictive health or target recognition.
- **Provocative:** The RAI Guidelines are not prescriptive. They are intended to provoke, surface questions, and spur discussions. The discussions sparked by the Guidelines are as important as the conclusions reached.
- **Useful:** The RAI Guidelines are intended to assist, not hinder project development. If not properly formed, there is a risk that ethics requirements may be viewed by vendors and project managers as obstacles to overcome, rather than essential components of successful AI development. The RAI Guidelines are designed to clarify roles and expectations, identify harms that can be avoided, and acknowledge unavoidable risks in any end product. Applied correctly, the RAI Guidelines will accelerate programs by commanding clarity of end goals, alignment of expectations, and acknowledgment of risks and trade-offs from the outset.

[5] Defense Innovation Unit. "Reducing the Time to Award: DIU's Commercial Solutions Opening." Webinar. June 4, 2020. https://www.diu.mil/latest/reducing-the-time-to-award-dius-commercial-solutions-opening

[6] Joy Boulamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research, 81 (2018):1-15. ********proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; National Transportation Safety Board. "Preliminary Report HWY18MH010." ***********.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

RESPONSIBLE AI GUIDELINES

DIU's RAI Guidelines aim to provide a clear, efficient process of inquiry for personnel involved in AI system development (e.g., program managers, commercial vendors, and government partners) to accomplish the following goals:

- ensure that the DoD's Ethical Principles for AI are integrated into the planning, development, and deployment phases of the AI system lifecycle;
- effectively examine, test, and validate that all programs and prototypes align with the DoD's Ethical Principles for AI;
- leverage a process that is reliable, replicable, and scalable across a wide variety of programs.

To this end, the DIU RAI Guidelines are framed within three major phases of the technical lifecycle of an AI system: planning, development, and deployment. Planning refers to the process of conceptualizing and designing an AI system to solve a given problem; development refers to the iterative process of writing and evaluating the computer code that comprises that system; and deployment refers to the process of using that system to solve the problem in practice.

> **"DIU has been able to better align its projects with the DoD AI Ethical Principles while driving the development of AI systems that are functionally superior and more rigorously evaluated."**

Below, we briefly describe the process laid out for each phase. DIU has also developed detailed worksheets, that instruct and guide AI vendors, DoD stakeholders, and DIU program managers on how to properly scope AI problem statements; these worksheets can be found in the appendix, and the most recent versions can be found at https://www.diu.mil/responsible-ai-guidelines

How to Follow the RAI Guidelines

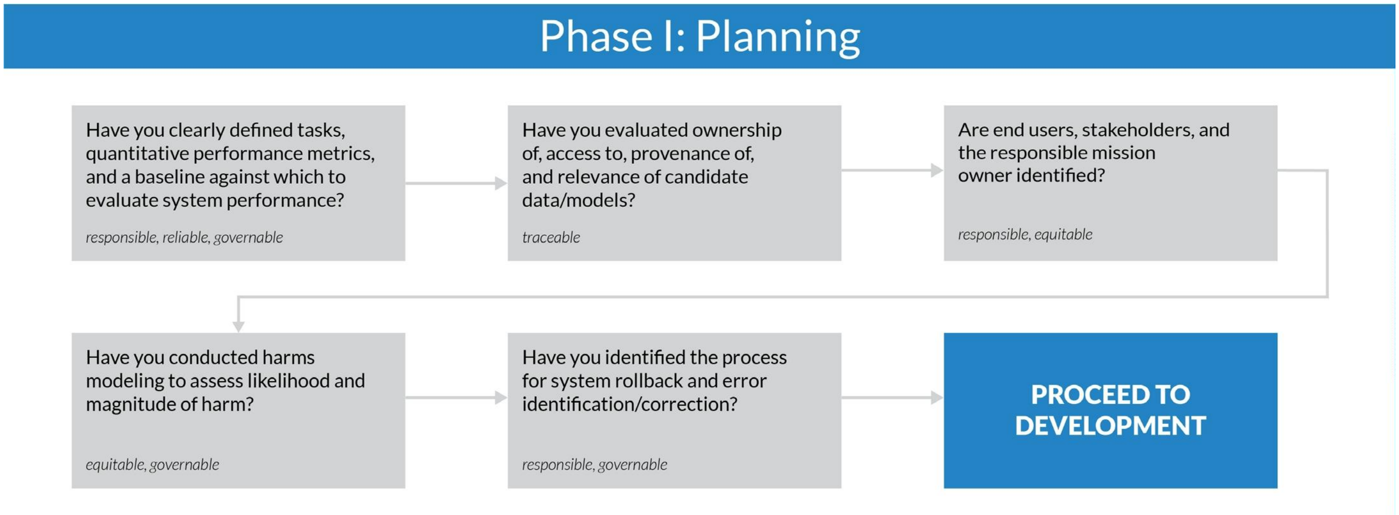
The following sections include graphical workflows that visualize specific considerations that AI vendors, DoD stakeholders, and program managers should consider (and address) before proceeding to the next phase of building an AI system.

Each workflow is supplemented by a more detailed worksheet (each included in the Appendix and matched to the planning, development, and deployment phases). The worksheets and workflows operate in concert and are designed to ensure that, for example, questions asked during the planning phase are sufficiently answered before stakeholders advance a project to the development and deployment phases.

This approach serves as both a documentation and verification mechanism; it ensures that projects that advance from planning to development and from development to deployment have met a rigorous vetting standard.

PLANNING PHASE

In the planning phase, personnel from the government agency requesting the AI system collaborate with the program manager to define its prospective functionality, the resources required to create it, and the operational context into which it will be deployed. Figure 1 presents the RAI Guidelines for the planning phase, which consist of five key lines of inquiry that are directly mapped to the DoD's Ethical Principles for AI.



Phase I: Planning

Have you clearly defined tasks, quantitative performance metrics, and a baseline against which to evaluate system performance?

responsible, reliable, governable

Have you evaluated ownership of, access to, provenance of, and relevance of candidate data/models?

traceable

Are end users, stakeholders, and the responsible mission owner identified?

responsible, equitable

Have you conducted harms modeling to assess likelihood and magnitude of harm?

equitable, governable

Have you identified the process for system rollback and error identification/correction?

responsible, governable

PROCEED TO DEVELOPMENT

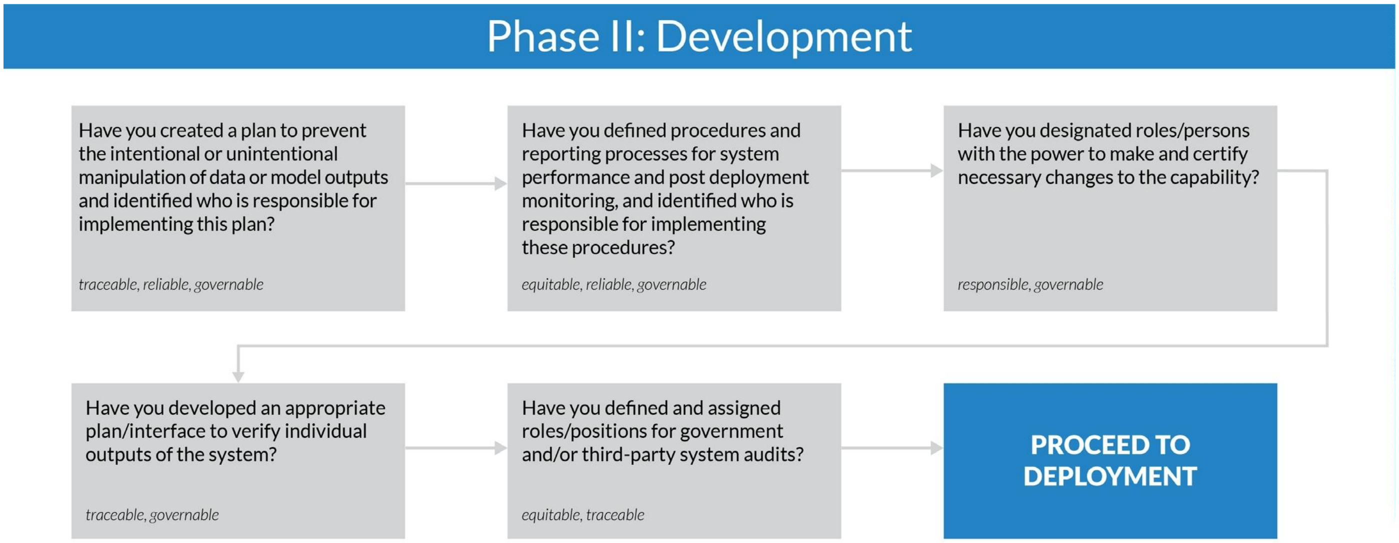
Figure 1: Planning Phase Workflow

The planning phase workflow ensures that AI is appropriate for the task and only applied after

- other methods (e.g., human-driven solutions) have been evaluated;
- success metrics and baselines are well-scoped;
- appropriate data (e.g. high quality, accurate, representative, etc.) is acquired to support the capability;
- stakeholders, mission owners, and end users (including potentially impacted populations) are appropriately considered and consulted;
- detailed risk assessments and harms modeling are conducted; and
- processes for reverting back from a malfunctioning system and identifying or addressing system errors are preemptively prescribed.

DEVELOPMENT PHASE

In the development phase, DoD and/or company personnel work to build out the planned AI system. Figure 2 presents the Responsible AI Guidelines for the development phase, which lays out five additional lines of inquiry that are directly mapped to the DoD's Ethical Principles for AI.



Phase II: Development

Have you created a plan to prevent the intentional or unintentional manipulation of data or model outputs and identified who is responsible for implementing this plan?

traceable, reliable, governable

Have you defined procedures and reporting processes for system performance and post deployment monitoring, and identified who is responsible for implementing these procedures?

equitable, reliable, governable

Have you designated roles/persons with the power to make and certify necessary changes to the capability?

responsible, governable

Have you developed an appropriate plan/interface to verify individual outputs of the system?

traceable, governable

Have you defined and assigned roles/positions for government and/or third-party system audits?

equitable, traceable

PROCEED TO DEPLOYMENT

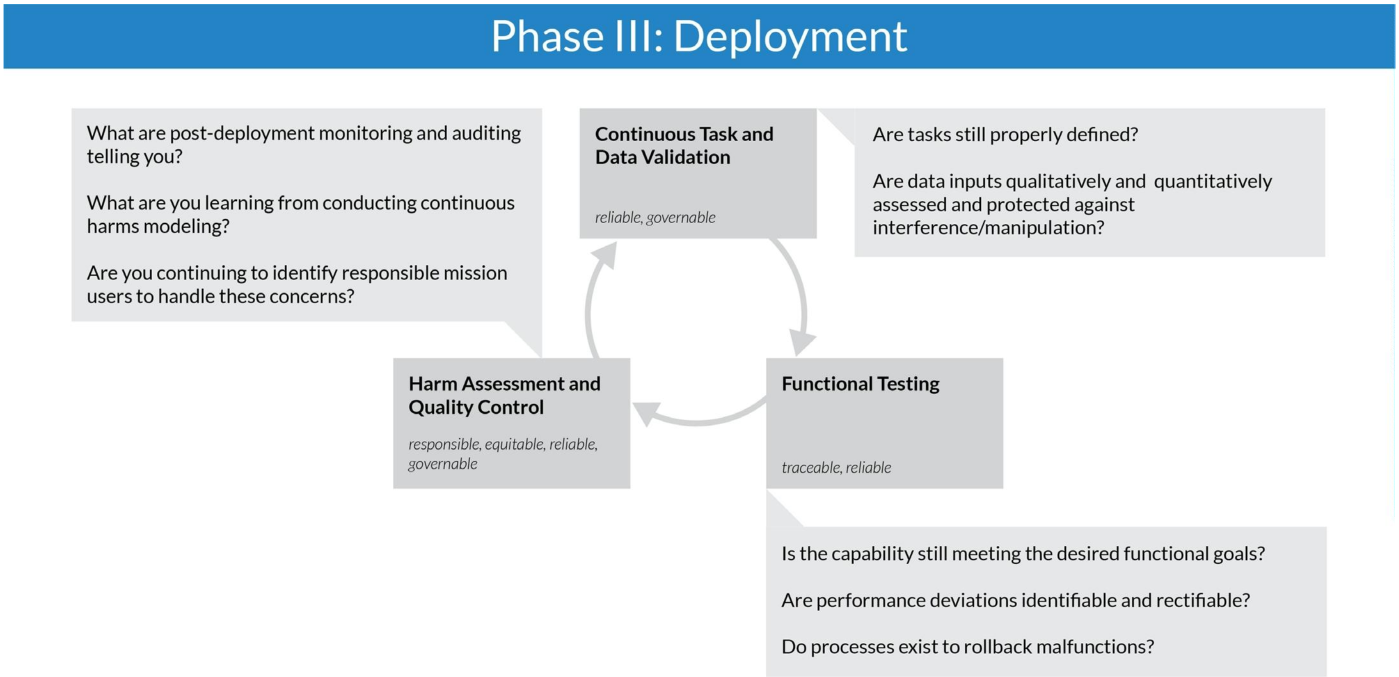
Figure 2: Development Phase Workflow

The development phase workflow focuses on

- mitigating the potential negative impact of data or model manipulation;
- delineating metrics and indicators for post-deployment monitoring;
- explicitly assigning authority to make changes to the capability;
- enabling users to understand how each system output is generated; and
- planning for routine system auditing.

DEPLOYMENT PHASE

In the deployment phase, DoD or company personnel make use of the AI system in an operational setting. Figure 3 presents the RAI Guidelines for the deployment phase, which describe concrete sets of continuous evaluation procedures that must be scoped and performed on an ongoing basis throughout an AI system's lifecycle.


Phase III: Deployment
Continuous Task and Data Validation
reliable, governable
Functional Testing
traceable, reliable
Harm Assessment and Quality Control
responsible, equitable, reliable, governable

What are post-deployment monitoring and auditing telling you?

What are you learning from conducting continuous harms modeling?

Are you continuing to identify responsible mission users to handle these concerns?

Are tasks still properly defined?

Are data inputs qualitatively and quantitatively assessed and protected against interference/manipulation?

Is the capability still meeting the desired functional goals?

Are performance deviations identifiable and rectifiable?

Do processes exist to rollback malfunctions?

Figure 3: Deployment Phase Workflow

The deployment phase workflow focuses on

- continuous task and data validation that ensures the original task specification and data inputs are still valid and secure;
- functional testing that evaluates whether the capability still performs the desired task sufficiently well to be operationally useful; and
- harms assessment and quality control to make certain that potential negative impacts on stakeholders are constantly reassessed and mitigated when necessary.

By assessing ongoing and upcoming projects in accordance with these Guidelines, DIU has not only been able to better align its projects with the DoD's Ethical Principles for AI, but also to drive the development of AI systems that are both functionally superior and more rigorously evaluated. In the case studies that follow, we describe several of these projects, how the RAI Guidelines were applied to them, and the outcomes we observed.

CASE STUDIES: RAI GUIDELINES IN PRACTICE

This section briefly summarizes the results of applying the Responsible AI Guidelines to two specific projects at DIU: Predictive Health and Countering Foreign Malign Influence. While the data types and objectives of these projects are very different, each demonstrates how the RAI Guidelines can provide value in practice by ensuring that AI capabilities are responsibly planned, developed, and deployed via a process that is both time and resource efficient.

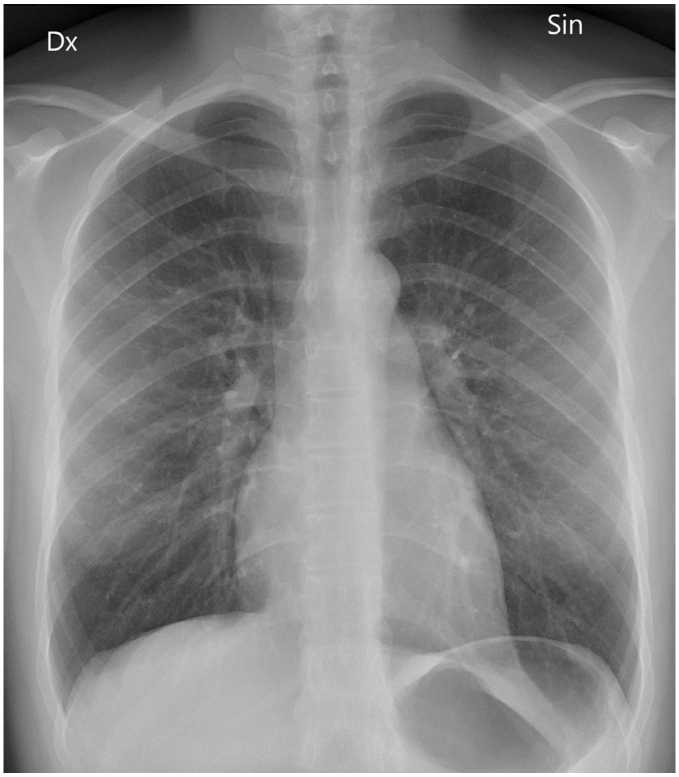
PREDICTIVE HEALTH

The project is a partnership between DIU, the Joint Artificial Intelligence Center, the Defense Health Agency, Google, Jenoptik, and Enlitic. The purpose of this project is to bring advanced machine learning capabilities for medical image analysis into military treatment facilities. The project has two functional branches: digital pathology and radiology. The radiology effort—executed in partnership with Enlitic—aims to incorporate algorithms that automate worklist prioritization and anomaly detection into clinical workflows for chest X-ray triage, head computerized tomography (CT) analysis, and lung nodule detection. The digital pathology effort aims to deploy machine learning models for detecting and classifying various types of cancer (developed by Google) in pathology slides to diagnostic Augmented Reality Microscopes developed by Jenoptik. This case study focuses on the radiology effort only and highlights the most important outputs from applying the RAI Guidelines at each phase of the AI lifecycle.

PLANNING PHASE

The planning exercise for the chest radiography branch of the Predictive Health program yielded several important conclusions that positively impacted the direction of the project.

First, as the team evaluated the task, metric, and benchmark, it became clear that the benchmark for this system is defined by real-world radiology workflows in which radiologists read images in the order they are received. Thus, if machine learning models could identify those cases most likely to require clinical intervention, turnaround time for treating cardiothoracic illness could be decreased. This provides a clear and useful quantitative metric—turnaround time for treating remarkable disease—that should be used to evaluate the system. Keeping this concrete metric in mind throughout the program helped the team both make actionable programmatic decisions and clearly define the value the system was intended to provide.



Chest radiograph[7]

Second, while evaluating the candidate data, it became clear that the standard Digital Imaging Communication in Medicine (DICOM) files associated with medical imaging data contain relevant metadata on scanning parameters, machine type, patient demographics, and other important pieces of information that can be used for downstream error analysis and harms modeling. Notably, without direct access to a small set of example data, it would have been difficult to make this useful determination conclusively.

Third, in addition to straightforward identification of radiologists as end users and patients as stakeholders, the team observed that population health professionals and various actors in and around the care environment should also be considered as stakeholders in this model development process. Thus, when evaluating model outputs, it is important to consider how large-scale use might affect both of these user groups. Without this explicit step in the planning workflow—which was intended to be provocative—the team might not have considered the AI system's effects beyond the immediate stakeholders.

The harms modeling step of the planning phase revealed that errors in this model could send individuals in need of critical care to the back of the chest radiograph reading queue. Domain shift[8] can be problematic for these models if, for instance, we

[7] Häggström, Mikael. Normal posteroanterior (PA) chest radiograph (X-ray). June 28, 2017. Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Normal_posteroanterior_(PA)_chest_radiograph_(X-ray).jpg

[8] Domain shift occurs when the data used to train an algorithm differs from the data encountered during deployment.

move deployment to a new patient population with a higher percentage of remarkable cases than was represented in the training distribution. Some edge cases can also cause errors. For instance, the model could accidentally be run on a different type of scan (e.g. a CT), and its output would be invalid. If the clinician is unaware of this error, worklists could be reordered in a counterproductive way.

To mitigate these potential harms, Enlitic explicitly tests models against multiple rare classes, on patients between 18 and 65 years old, and on all U.S. cases for U.S. models. Their testing data generally comes from the last ten years. Enlitic typically retrains or re-evaluates a model if the class balance changes by more than two standard deviations. Note that participation of both the government team, the vendor, and the end users—in addition to review of the relevant academic literature—was extremely important to this exercise, as each group had specific knowledge that contributed to potential harms being more fully and completely identified. Many of the items identified during harms modeling were directly integrated into the Testing and Evaluation (T&E) plan for the program in order to realistically mitigate unfavorable outcomes.

Finally, while the process for system rollback is relatively straightforward in this case—the radiologists would simply return to their pre-capability workflow—a rigorous process orchestrated by the Food and Drug Administration (FDA) is required not only to certify model performance, but also to handle any errors observed in practice. Spending time understanding this process was highly valuable to the project team, as it helped not only to define how rollback was expected to be handled for this particular system, but also provided insight on best practices from other public and private sector entities.

DEVELOPMENT PHASE

In the development phase, it became clear that the processes in place for FDA approval directly incorporated several of the focus areas emphasized by DIU's RAI Guidelines. These included: procedures for system performance measurement, post-deployment monitoring, individual system output verification, and model updates. For instance, the company maintains a traceability matrix that keeps track of model versions and verification and validation results. If an error is identified in practice by a clinician user, a corrective action procedure (CAP) is initiated: a root cause analysis is performed, and engineers determine a solution and distribute a new standard operating procedure along with new model versions as required. The FDA is notified about any changes to the model, and a new set of weights are deployed to multiple customers via a parameter server. Implementation of procedures necessary for maintaining patient privacy were considered sufficient for preventing data or model manipulation. Through this development process, DIU realized that our RAI Guidelines can complement other Ethical AI efforts that are ongoing across the U.S. government—as in the case of the FDA—while providing a framework that other organizations can replicate and adapt for their own needs.

DEPLOYMENT PHASE

Discussions around deployment yielded several useful pieces of information that should inform long-term system use. Data validation—i.e. ensuring that data provided to the algorithm is within appropriate parameters—is, at present, performed by the user. This means that the user must be sufficiently trained to responsibly use this capability in practice, and cannot rely on automated checks to ensure correct application of the models. Continuous functional testing and harms assessment processes do exist; these aspects of deployment are primarily handled through the performance reporting process described above that is implemented under the auspices of the FDA. Similar to the development phase, the deployment phase of this project has leveraged a combination of public and private sector expertise.

OUTCOMES

This project is ongoing. Important outcomes to date include several improvements to sections of the T&E plan that incorporate clear baselines from academic literature against which algorithmic performance may be tested, explicit delineation of demographic and disease-based subpopulations upon which model degradation should be evaluated, and evaluation of the value added to clinical workflows. Further, the project team was able to learn about and ultimately leverage best practices from both government and industry, which will improve the team's ability to manage AI projects in the future.

COUNTERING FOREIGN MALIGN INFLUENCE

The Countering Foreign Malign Influence project is a partnership between DIU, DoD, and Quantifind. The project aims to better support DoD analysts by leveraging analytics derived from commercially available information and publicly available information to identify, track, and counter transnational criminal groups attempting to mask their identities and activities. It primarily makes use of open source data to support construction of knowledge graphs that allow for more efficient use of analyst time and surface relationships between entities that would be difficult for human analysts to identify due to the large volume and complexity of data that must be analyzed.

PLANNING PHASE

As the team began to evaluate the task, metrics, benchmark, and candidate data for this project, it became clear that this work would require a substantial amount of engagement between Quantifind and the DoD project team. Building and using a knowledge graph involves a large number of different modeling and analysis steps including named entity recognition, relation extraction, visualization, and risk model application. Thus, the RAI Guidelines would need to be applied in a manner that was both realistic and useful—that is, sufficiently detailed to address each of these modeling steps without hindering the overall development process.

In this context, the team chose to focus on two specific types of documentation during their planning phase: model cards, initially put forth by Mitchell et al., as a way to document machine learning model construction and analysis, and risk cards, a type of document Quantifind has created to record and explain design choices and assumptions.[9]

As part of these adaptive efforts to implement the DIU RAI Guidelines, the team defined clear metrics (and ontologies where necessary) for each task (e.g. named entity recognition, relation extraction, etc.), described each data source and its potential strengths and weaknesses in relation to the task, and considered how to most effectively and responsibly integrate structured and unstructured datasets across multiple languages, in multiple formats, and from a wide variety of sources (news media, commercial data registers, and a variety of data aggregators). A major takeaway from these first two sections of the planning phase was that complex machine learning systems that leverage many interconnected subcomponents need to not only be analyzed from an end-to-end perspective, but also broken down into their constituent parts such that each component can be planned individually.

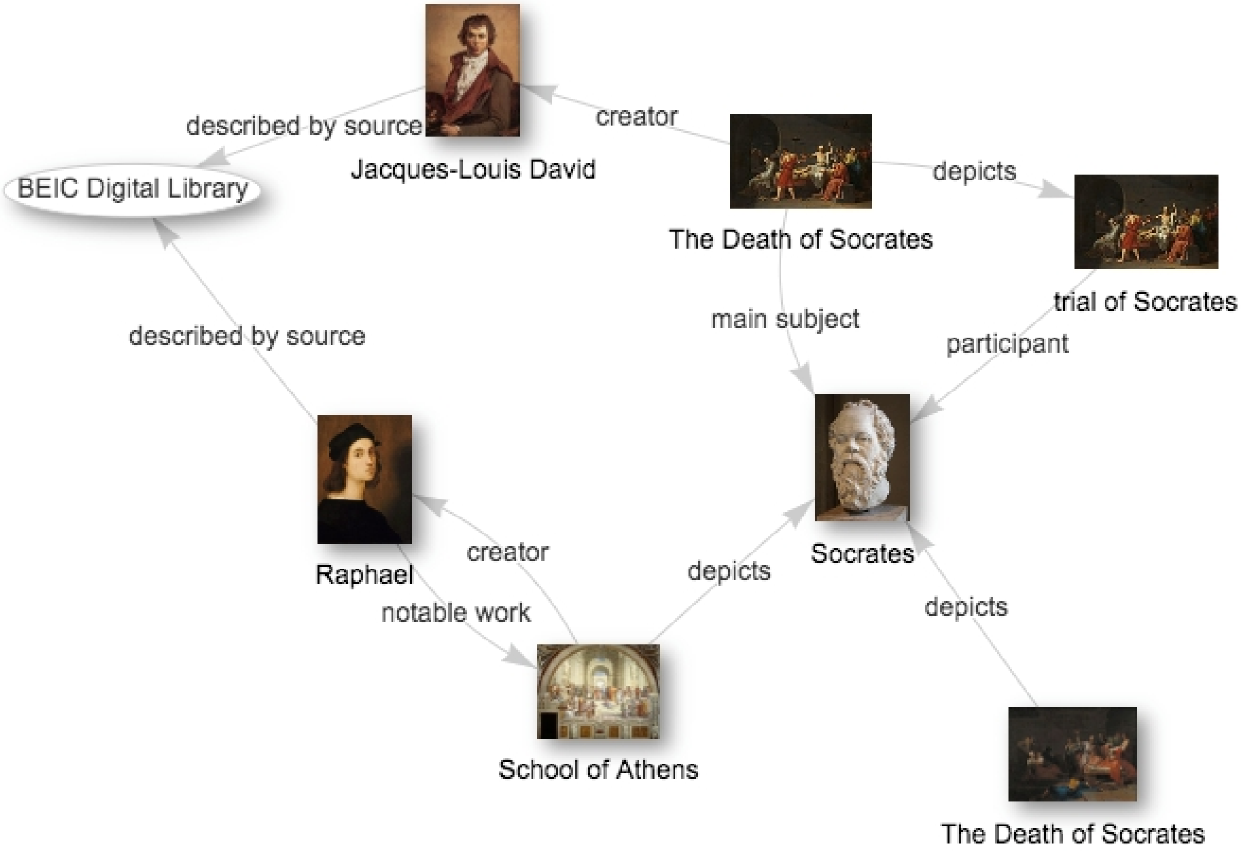
Through the provocative process of identifying end users and stakeholders and performing subsequent harms modeling, the team identified a number of nuanced issues that would be important to address during development of the capability. One compelling example stems from the fact that Quantifind must combine relation extraction across content types (structured and unstructured) to determine how an entity (person, company, etc.) is related to other entities and activities within the same text. The context of the situation being described is a significant factor in this determination and that nuance needs to be reflected in model outputs. It is crucial, for instance, to distinguish between an undercover journalist and genuine members of criminal groups even if observable activity is identical. The Quantifind team planned to address the potential harms to individuals like undercover journalists—who were identified as stakeholders in, but not end users of their system—in several ways.

First, Quantifind aimed to construct models in such a way that data and context relevant to this source of potential harms were provided as inputs to the model. Second, Quantifind worked to include raw data such as the source documents that heavily influenced a given model result as part of the system output. The first intervention improves the chances that the model will be able to address this particular source of harm, while the second serves as a human-in-the-loop check that allows an analyst to leverage the model to identify relevant source documents, but ultimately make a determination using human reasoning. These activities established additional mechanisms for measuring and quantifying platform performance on nuanced subtasks.

In this case, the process for system rollback was relatively straightforward: analysts would revert back to the workflow they currently use in practice. Importantly, this means that analysts must still be trained to perform the analytic tasks that Quantifind's platform is meant to support without leveraging the capability, which is an important nuance for long-term planning and resourcing on the part of the DoD end user.

It is worth noting that throughout the planning phase, the project team used both model cards and risk cards to support technical development and communication with government partners. These model cards contain a substantial proportion of the information that the planning phase is intended to elicit. Model cards are referenced in the planning worksheets, and provide information about each model in the AI system. John Stockton, a co-founder of Quantifind, describes them as explaining "what goes in, what goes out, what [the models] do, with specific examples." Stockton emphasizes the importance of people in the field knowing what their tools are and are not good at, and understanding how to fit an AI system into an overall workflow. Importantly, these model card efforts are a consultative process with subject matter experts and technical

[9] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. "Model Cards for Model Reporting." Paper presented at the *Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA, USA*. 220-229. https://arxiv.org/abs/1810.03993



Example knowledge graph describing relationships (e.g. depicts, creator, described by, etc.) between entities (Socrates, Raphael, etc.) in the domain of art history. Entities in a graph tend to be referred to as "nodes" while relationships are often referred to as "edges." [10]

experts collaborating to make a system that enhances analysts' work. In this particular application, it was critical for Quantifind to collaborate with domain experts and data scientists to make standardized, operational definitions that enable abstract concepts in foreign malign influence to be expressed as concrete machine learning tasks with clearly identifiable performance metrics, data requirements, stakeholders, end users, and harms modeling emphases.

DEVELOPMENT PHASE

During development, Quantifind continuously tests and monitors their models to identify changes over time; this simultaneously increases the probability of building an effective capability and provides the necessary building blocks for post-deployment monitoring. Findings from these iterative development processes—as well as continued analysis of how adversarial actors could seek to undermine the capabilities being developed—are continuously integrated into the model cards and risk cards alongside advancements in the underlying modeling approach (e.g. leveraging embeddings from cutting edge natural language processing models).

Operational risks identified during the development process range from explaining the nuance required to decide which topics should be considered risk indicators to evaluating the ability of named entity recognition and relation extraction models (trained on historical data) to continue to perform well once deployed. As Quantifind continues through the development process, the updates to model cards and risk cards will ensure that, when the system is completed, the performance envelope will be clearly established, design choices and assumptions will have been made clear for users, and operators will be able to make decisions about where and when it would be most appropriate to deploy the capability in their operational context.

Furthermore, to create a process by which individual outputs of the system can be verified and evaluated, the Quantifind team conducts user experience studies to determine appropriateness

[10] Wikidata knowledge graph example using SPARQL .October 15, 2019. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Wikidata-knowledge-graph-fuzheado-metobjects-2.png

of use and understand how the user interface influences trust and reduces the probability of misinterpretation. This involves obtaining direct feedback on the platform on a regular basis, exposing source documents that support any given extracted relation, and experimenting with different approaches to display model confidence to end users. Importantly, while this process is critical to building DoD confidence in the capability, it also provides valuable feedback that Quantifind can use to improve its commercial offerings.

Finally, given that the capability is still in development, the team and DoD partner are continuing to discuss who would have authority to make changes to the system in practice, and how system auditing will take place in the future. It is important to note that because development is iterative, it may sometimes be necessary to update assumptions about the responsible authority and auditing approach as the form factor and intended deployment pattern of the capability becomes more concrete.

DEPLOYMENT PHASE

While this project has not yet entered the deployment phase, the model cards, risk cards, and other pieces of documentation have set the project up well for these conversations. The team expects that the material contained in these pieces of documentation and the continuous testing protocols refined during the development phase will enable clear delineation of what needs to be accomplished to perform continuous task and data validation, functional testing, and harms assessment and quality control. Deployment phase documentation will be created as the project approaches this phase of maturity.

OUTCOMES

The Countering Foreign Malign Influence project is progressing through the development guidelines and the project team is beginning to consider the deployment activities as they move from prototyping into production. The team continues to work closely with their government counterparts and to integrate helpful new practices into their work. The documentation described above has been an extremely important outcome for this project, as it has not only documented the work Quantifind has done to a degree that will be important in enabling real-world adoption, but it has also identified important issues that should be addressed before the system is deployed. Going forward, for instance, it will be particularly important to analyze the concrete tradeoff between performance gains obtained by leveraging larger language models and the potential biases or performance irregularities that such models can introduce, as well as to continually measure performance at both individual model and end-to-end system levels.[11]

A second valuable outcome of this project was the realization that DIU's RAI Guidelines mirrored many of the processes that Quantifind uses internally to integrate Ethical AI principles into its work. Quantifind described the question-response style of the planning worksheet as providing the "opportunity to proactively communicate to the government the existing company processes, standards, and known problem areas to get them on the table," and stated that integrating DIU's RAI Guidelines into their process helped to "mediate a good two-way dialogue" on RAI. This two-way dialogue benefits both public and private participants and creates a model for collaboration that can be replicated across the U.S. Department of Defense.

[11] Emily M. Bender, Angelina McMillan-Major, Timnit Gebru, Margaret Mitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Paper presented at the *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* 610-623. ACM, New York, NY, USA. https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

LESSONS LEARNED

This section lays out key lessons DIU learned while implementing the RAI Guidelines.

Spend time on metrics: *How you measure success determines success.*

Machine learning algorithms seek to optimize pre-defined metrics. Selecting these metrics, and ensuring they align with operational goals, is crucial. In the field of AI safety this is called the alignment problem—the challenge of making sure that AI outcomes align with human goals.[12] For example, in the case of predictive health, metrics for assessing models could include accuracy and weighted F1 score, while the operational metric could be turnaround time for treating cardiothoracic illness.

Many of the most discussed ethical issues in AI—such as bias in facial recognition—can arise because an algorithm is optimizing the wrong metric. The default to accuracy or other high-level metrics means that algorithms often perform poorly on underrepresented groups[13] or edge cases that may be infrequent but incredibly consequential. The decision of what metric to use is, oftentimes, of ethical consequence. For example, a hiring system might be optimized to ensure that candidates of different backgrounds are all treated according to the same standard or that candidates of different backgrounds are equally likely to be selected. While these two goals both align with intuitive notions of fairness, they are often incompatible in practice.[14] Consequently, it is important to have explicit alignment around what definition of fairness—and similar ethical concepts—will be adopted for the purposes of a given project.[15]

Account for technology/task fit: *High-risk applications should be paired with low-risk technology, and vice versa.*

Many applications where AI is used commercially—such as music recommendation systems—are inherently low risk: the worst case scenario is that a user becomes frustrated or selects an option that was not optimal for them. However, errors made by AI applications in national security use cases can carry much higher costs.

In order to continuously develop better AI-enabled systems that perform effectively in the field, it is critical to calibrate technical and operational risks appropriately. The correlation between technical and operational risks should be negative. Low-risk operational applications where the cost of an error is limited—such as optimizing aircraft cleaning schedules—are excellent testbeds for high-risk technologies. On the other hand, high-risk operational applications, where even the smallest error could be disastrous, should either make use of low-risk technologies with proven track records or require human oversight over every model output built into the proposed AI-enabled workflow.

A good example of defining an AI-enabled workflow for a high-risk system occurred in the Predictive Health project that is included as a case study in this paper. Because the potential downside of misdiagnosis on medical imaging is high, humans will ultimately review every image that has undergone analysis using the AI system to confirm the X-ray results. Thus, while the AI model can make diagnosis more efficient by directing physicians to abnormal X-rays more rapidly than would their usual workflow, erroneous model output would not cause degradation in the end-to-end diagnosis process.

> **"In order to continuously develop better AI-enabled systems that perform effectively in the field, it is critical to calibrate technical and operational risks appropriately."**

There are some national security contexts that may require the pairing of high-risk technologies with high-risk applications. This need is more likely in situations where an AI-enabled system is required to address an urgent national security crisis event where, for instance, the failure to detect, alert, or respond to an adversarial threat results in an immediate negative outcome (such as loss of life). In these situations, the application of high-risk technologies to high-risk mission areas may be warranted and should be considered. However, mission users, policy makers, and decision makers should be advised of the potential consequences when a high-risk technology is fielded to address a high-risk mission area or need.

Incorporate industry best practices where appropriate: *DoD organizations can learn from commercial sector advances in ethical AI development.*

Many of the AI solutions sought by the DoD have parallel applications in the private sector, where commercial organizations are independently developing ethical frameworks to guide and inform their AI development activities.[16] In project execution, DIU has often found that vendors' internal processes

[12] Brian Christian. The Alignment Problem: Machine Learning and Human Values. (New York: W. W. Norton & Company; 2020).
[13] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification in Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR 81:77-91, 2018. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline
[14] Kailash Karthik Saravanakumar. The Impossibility Theorem of Machine Fairness A Causal Perspective. Cornell University. July 12, 2020. https://arxiv.org/abs/2007.06024
[15] Bryce Goodman. Hard Choices and Hard Limits for Artificial Intelligence. Cornell University. May 4, 2021. https://arxiv.org/abs/2105.07852
[16] Algorithm Watch. "AI Ethics Guidelines Global Inventory." https://inventory.algorithmwatch.org/

mirror the activities included in the RAI Guidelines. For example, Quantifind—the commercial partner in the Countering Foreign Malign Influence project—already had a robust ethical framework that they employ while developing commercial applications. This convergence opens a space for the cross-pollination of best practices that can benefit both the government and its commercial partners in the future. Such interaction also narrows the gap between the DoD's own standards for ethical AI development and those already in practice in the commercial and private sector, which can substantially accelerate responsible adoption of AI systems by the DoD.

Set customer and vendor expectations: *AI is not magic, and developers should not act like magicians.*

Like any technology, AI has both its benefits and its limitations. A clear-eyed assessment of the issues laid out in the RAI Guidelines, accompanied by a realistic estimation of long-term resourcing and sustainment requirements, will enable pragmatic decision making about the potential costs, benefits, and risks of developing an AI-based system for a given application. Importantly, many systems that are prototyped for national security use cases may simply not make the cut for responsible deployment; deciding not to further pursue an AI capability should be an acceptable outcome of any potential project. Program managers should be clear about responsible AI expectations from day one; vendors should be encouraged to come forward with performance issues, and DoD partners should remember that any flaw in an AI system represent an operational risk to DoD personnel.

Invest time and resources in documentation: *AI capabilities cannot be used confidently without comprehensive, even-handed documentation.*

The process of working through the RAI Guidelines on real projects has reinforced the importance of precise, descriptive documentation to the viability and efficacy of AI capabilities for the DoD. In order for a user to be confident in leveraging a capability for a given mission, they must not only be convinced that the system works as intended, but also that the user's problem corresponds exactly to what the AI system was built to solve. In other words, users must be convinced that model testing, evaluation, verification, and validation has been performed for their operational context. Because this is difficult to achieve for every possible user, AI capability documentation—such as data cards, model cards, test and evaluation plans, auditing results, etc.—is absolutely critical to enabling these capabilities to be adopted efficiently and effectively. As a result, program managers and senior leaders must consciously budget both resources and time to create these sets of documentation for AI capabilities. Without this investment, the DoD will end up with a large number of AI science projects that do not ultimately provide value to the operator.

CONCLUSION

Having spent over a year developing, testing, and iterating upon the DIU RAI Guidelines with partners from industry, academia, government, and civil society, the DIU team believes that the current version of the Guidelines represents a useful starting point for defining a process by which the DoD's Ethical Principles for AI can be operationalized on acquisition programs. We emphasize that this work is not intended to be a final product, and we actively seek advice, feedback, and constructive criticism. The RAI Guidelines and any derivative or complementary material will be routinely updated, reevaluated, and iterated upon to maximize the degree to which DIU AI capabilities are developed in alignment with its foundational ethical principles.

APPENDIX

RAI GUIDELINES: WORKSHEETS & COMMENTARY

The following tools are designed to help AI companies, DoD stakeholders, and DIU program managers implement RAI Guidelines when developing AI applications for DoD end users. The planning worksheet is primarily intended to be a collaborative effort between DoD (or other government) stakeholders and DIU program managers; the development and deployment worksheets should be a joint effort between the commercial company/ies on contract, DoD stakeholders, and DIU project team members. These tools are not exclusive to DIU and others may apply or adapt them as befits their needs.

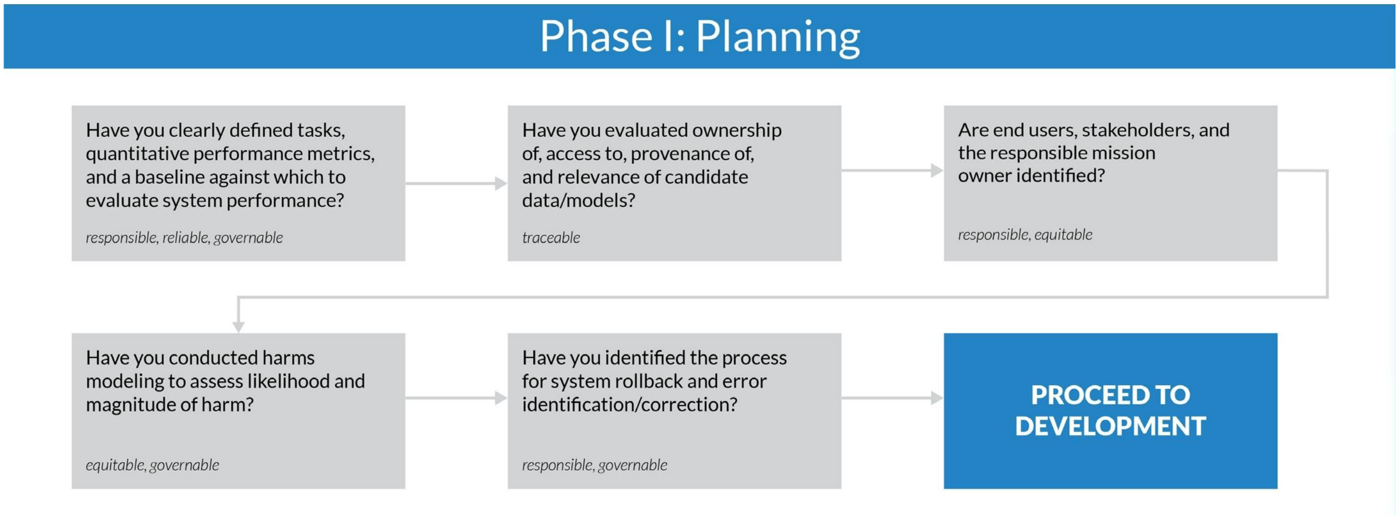
Please note that the materials included in this Appendix are current as of the publication of this paper, November 10, 2021; fillable PDF versions of the following worksheets and future iterations of both the worksheets and commentary will be accessible on DIU's website at https://www.diu.mil/responsible-ai-guidelines.

Phase 1: Planning

The planning worksheet is to be completed by the government agency requesting the AI system with the program manager prior to awarding a prototype agreement, and then updated as needed once the commercial vendor is selected.

Directions: Respond to the following questions in the order they are presented and include notes about your conversation(s) with regard to applicability for planning efforts. Include descriptions of what has already been completed and what work is left to be done (if applicable). Include context as appropriate, such as a timeline for completion and current status. Please provide a justification if the question is not applicable to the project or the issues raised will be resolved at a later date.

PLANNING WORKSHEET // PLANNING PROCESS FLOW



Phase I: Planning

Have you clearly defined tasks, quantitative performance metrics, and a baseline against which to evaluate system performance?

responsible, reliable, governable

Have you evaluated ownership of, access to, provenance of, and relevance of candidate data/models?

traceable

Are end users, stakeholders, and the responsible mission owner identified?

responsible, equitable

Have you conducted harms modeling to assess likelihood and magnitude of harm?

equitable, governable

Have you identified the process for system rollback and error identification/correction?

responsible, governable

PROCEED TO DEVELOPMENT

QUESTIONS

- Have you clearly defined tasks, quantitative performance metrics, and a baseline against which to evaluate system performance? *Define these elements.*
- Have you evaluated ownership of, access to, provenance of, and relevance of candidate data/models? *Identify these relationships.*
- Are end users, stakeholders, and the responsible mission owner identified? *Identify these groups or individuals.*
- Have you conducted harms modeling to assess likelihood and magnitude of harm? *Explain your approach. What did you find?*
- Have you identified the process for system rollback and error identification/correction? *Define your process.*

COMMENTARY ON PLANNING WORKSHEET

The following is intended to accompany the questions for the planning flow and provides additional context for the questions to guide you through the evaluation process.

1. Have you clearly defined tasks, quantitative performance metrics, and a baseline against which to evaluate system performance?

CLEARLY DEFINED TASKS FOR AI SYSTEMS

By AI system, we mean a computer system related to the development, testing, management, delivery, and/or research of machine learning, statistical decision-making, and advanced analytics.[9] By task, we mean the intended function of the capability—i.e.,what the capability will enable a human or another system to do.

The first question to ask in any AI project is whether AI technology provides a unique, non-marginal benefit, or whether an alternative method should be selected. An AI approach may be advantageous if the task involves[10]

- natural language processing: e.g. translating text from one language to another, generating a summary;
- recognizing a pattern or object: e.g. detecting whether a transaction is fraudulent[11], or identifying types of vehicles in images;
- personalization/customization: e.g. recommending relevant documents based on past search history;
- detection of low occurrence events that change over time: e.g. identifying which parts are likely to break; or
- predicting future events: e.g. forecasting weather events.

On the other hand, AI is generally not the optimal approach if the task requires

- complete predictability: e.g. knowing how the system is likely to react to future events;
- complete transparency, interpretability, and explainability: e.g. knowing exactly how and why the system recommends or take a particular action;
- complete assurance: e.g. where a single error could be extremely costly;
- subjective judgement: e.g. where different people would reasonably disagree about the best outcome; or
- solving existing human problems: e.g. clarifying an existing process that is confusing and/or problematic; or fixing existing problems in sets of data (such as bias).

If you decide that AI provides the best approach, a clearly defined task will require a description of what the system will enable or accomplish.

> *Example of a well-defined task:*
> Localize and classify building damage from pre- and post-disaster satellite imagery using computer vision in order to achieve a performance threshold of 90% (defined by intersection-over-union for the detections vs. ground truth) and reduce turnaround time by 20% for imagery analysts.
>
> *Example of a poorly defined task:*
> Improve damage assessment using machine learning.

The first task is well defined because the purpose of the system (localize and classify building damage), the primary *end users* (post-disaster analysts), the *input* (satellite imagery), the *output* (damage assessments), and both quantitative algorithmic and operational metrics for evaluation (e.g. intersection-over-union for detections vs. ground truth and turnaround time) are all identified.

The second task is not well defined because the purpose, the end users, and the inputs and outputs are not identified. One must always be wary of AI projects that seek to "improve" performance—you should always ask, "What aspect of the desired solution is to be improved: speed? breadth of information?" There are often trade-offs between speed, precision, recall, explainability, etc., so a clearly defined task is critical for guiding technical decisions down the line.

[9] For the purposes of this document, AI refers to machine learning (ML). This is distinct from the definition offered by the Defense Innovation Board. Defense Innovation Board. "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense." https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

[10] Google, "User Needs + Defining Success." in People + AI Guidebook, https://pair.withgoogle.com/chapter/user-needs/

[11] NB. If fraud is not well defined, systems that automate fraud detection are likely to perform poorly.

Potential questions for program managers/vendors:

- Is AI suitable for the task at hand?
- What is the specific task that the system performs?
- What is the system input and output required to perform that task?

QUANTITATIVE METRICS

Machine learning is math. If we cannot state our objective mathematically, we cannot use machine learning. Consequently, clear and consistent metrics that define how and when a model is successful are critical for implementing and using AI models. In the majority of cases, the lack of a quantifiable metric is a non-starter for training and using AI methods. Common metrics include accuracy, logarithmic loss, mean squared error, f1 score, intersection-over-union, perplexity, etc.

AI models are guided by mathematical formulas or reward functions that determine the success or failure of the AI system. A properly designed reward function will account for trade-offs in precision and recall, and is built with the end-user experience in mind. An AI system that is optimized solely for precision will only return results that are perfectly matched with user preferences. Such a model excludes potentially relevant results that are unknown to the user and can ultimately limit or hinder the overall user experience. Assessing reward functions for such trade-offs ensures that an AI system produces results that are inclusive and properly calibrated to minimize potentially negative outcomes over time.

Deliberate care must be taken to ensure appropriate metrics are utilized for a given task. Consider accuracy, a common metric for measuring the performance of an AI system. Assume that a dataset consists of 900 apples and 100 oranges. Accuracy, as traditionally computed, captures performance with each category treated equally. Creating a simple program that always returns the category of apple will be 90% accurate, even when presented with clear examples of the orange class. If we are interested in identifying a low-probability event, or have a dataset with unequal frequencies in different categories, accuracy is often not an appropriate metric.

Potential questions for program managers/vendors:

- Which metrics will you use to measure system performance?
- Why are those the correct metrics?
- What situations would lead to the metrics being optimized without the intended result being obtained?
- Might the metrics need to change as the system behavior changes in response to deployment? (considering feedback loops)
- How accurately or precisely can performance on each task be measured?

BASELINE

A baseline is a measure that allows for a comparison of performance on the task of interest before, during, and after a project. In simple terms, a baseline will let us know whether our AI system is worse, equal to, or better than the *status quo*. This can be the basis for both qualitative and quantitative assessments of how well the system is performing.

To establish a baseline, we begin by asking how the current task is being performed. If it is done manually, we must conduct a user study to standardize and quantify the quality of manual performance, the inter-rater reliability,[12] how long it takes for the task to be completed, and any other factors (e.g. importance of the ability to seek redress, transparency, due process, etc.). It is especially helpful if performance can be captured quantitatively. In the cases where it cannot, having a well-defined qualitative baseline (e.g. higher confidence in decision making, access to more accurate information, etc.) can still provide a basis for comparing the AI system.

There are a number of reasons why establishing a baseline *prior* to system development is important:

1. It allows project owners to prioritize what really matters for success. Sometimes there may be a trade-off between accuracy and speed. Knowing the baseline will allow the team to navigate these trade-offs by understanding what is currently acceptable and/or in deployment.
2. It provides an indication of minimum acceptable conditions for project success. A baseline allows project owners to assess the quality of predictions at various thresholds of the defined metric. In one case an 80% accurate solution may provide usable results for a team, whereas, in another case, useful results require 99% accuracy. Such thresholds should be documented at the outset of the program, and end users should be aware of the limitations of the system, which should be quantified via explicit metrics.
3. It enables a constant comparison during deployment. If we have established a performance baseline we can continually compare whether we are meeting that baseline. If not, we may need to modify or rollback the system.

[12] Inter-rater reliability is a measure of the degree of agreement between multiple data annotators.

Potential questions for program managers/vendors:

- How is the task currently performed?
- What is an acceptable minimum performance threshold?
- What are the most important evaluation criteria (e.g. speed, volume of data processed, quality of output, etc.)?

2. Have you evaluated ownership of, access to, provenance of, and relevance of candidate data/models?

OWNERSHIP

When considering the role that data plays in building AI systems, one must consider the data collection process, the data ontology, the data quality controls (including transformations), and finally, the data itself. It is critical that ownership of, and responsibility for, all components of the data pipeline used to train models is clearly specified and understood by all parties involved. The same is true of the models produced by training. Both may have implications for the cost structure and the government's ability to improve the system in the long term (e.g. by avoiding vendor lock-in).

ACCESS

Access to data and the data pipeline are critical before, during, and after the duration of the program. Usage rights, permissions, classification concerns, and distribution protocols should all be identified and contractually documented. Vendor lock-in is likely to occur if the data and the data pipeline are inaccessible due to proprietary data formats and protocols.

PROVENANCE

It is not sufficient to have data. One must know from where the data was sourced (e.g. what sensor); what transformations have been applied to it; and who labeled it, when, and for which task. Understanding where the data came from, how it was transformed, and how it was modified during the training process is critical. Each phase of the data pipeline introduces new biases and potential sources of fragility in the data. For example, only using imagery that is sourced from California to train a wildfire detection algorithm may limit the ability of that algorithm to generalize to other geographies. More data is not always better: training an algorithm on historical data may recreate historical biases around sex or race.

One should always ask:

- Why was this data collected?
- How was this data collected?
- Where was this data collected?
- Who did the collecting?
- Who organized the results?

RELEVANCE & REPRESENTATIVENESS

The data accessed must be of high quality in the sense that it is relevant to the task at hand. Common reasons for low data quality include insufficient dataset size, poor-quality labels, suboptimal definition of output schema (e.g. the classes in a classification problem are not operationally meaningful), or bias that will result in poor operational performance and/or unacceptable discriminatory outcomes.

Relevant data can seem counterintuitive at times. For example, if performing a task that involves detecting buildings, a dataset may be collected that includes buildings of different types. However, it is critical to include examples of imagery that have no buildings at all! Without these negative examples, an AI model may operate in an undefined space when given an example of a field, resulting in many false positives.

Potential questions for program managers/vendors:

- Who will own and manage the dataset and models?
- What format is required for input data?
- How will the provenance of the data be documented and shared with end users?
- How was the dataset collected, constructed, produced, and curated?
- Is the data relevant (current) and operational?
- What is a sufficient dataset size, and how was this determined?
- Who will label the data? how reliable are the data labels? how will labels be managed?

- What is the distribution of classes within the dataset (or equivalent for non-classification problems)?
- What checks have been completed to assess for bias and ensure representativeness?
- Does the intended use of the dataset align with the manner in which it was collected?
- Does the dataset contain personally identifiable information (PII)? Will the dataset be combined with other datasets that may then reveal PII? If so, what precautions will be taken to protect the privacy and welfare of data subjects?
- What steps will be taken to ensure the data is appropriately secure during and after the project?
- Have you considered/will you be using a documentation tool like model cards or datasheets for datasets? If not, why not?
- Are end users, stakeholders, and the responsible mission owner identified?

3. Are end users, stakeholders, and the responsible mission owner identified?

END USERS

AI does not function on its own—it is part of a human-machine system. For AI to be useful, it must address a user's needs. The user experience is as important (if not more) than algorithm performance. Users should be consulted extensively in the planning phase to ensure that the machine learning task matches their needs.

Identifying end users involves asking, "Who are the people (end users) that will be the primary user of the system?" Be as specific as possible with regard to role, responsibility, needs, etc. It is advisable to conduct user need analysis with end users (or close proxies) at an early (prototype) phase to ensure a match with user needs.

It is important to first examine existing workflows to deter mine how the end user currently accomplishes the tasks they seek to optimize with the AI solution. By examining their current process, development and design teams will better identify which aspects can be meaningfully enhanced by an AI solution versus those that would not benefit from its incorporation or would be degraded by it.[13] In some cases, a simpler, rule-based solution may be more appropriate.

Rule-based solutions are easier to build, develop, and maintain when compared to their AI counterparts and should be explored as potential alternatives as part of the user engagement process. Planning teams should consult a broad sampling of the potential user base to determine if and when an AI solution is best applied.

One important question to address is whether the AI will be *automating* a task or augmenting a user's ability to perform that task. The decision of when to automate vs. augment will hinge on a number of factors, such as:

- How consistent is the task? If the answer is very, it may be easier to automate. If it is more variable, it is likely that augmenting a user's abilities is a better route.
- How important is it that an individual can be held responsible for outcomes? If the answer is very, augmentation is preferred.

This automation vs. augmentation guide can help assess tasks that are best delegated to an AI system versus those that benefit most from the addition of AI as a supplement to the end user. The combination of automation and augmentation should simplify and improve the eventual output of the AI system while meeting the needs of the desired end users.

STAKEHOLDERS

Stakeholders are people who are either using the system outputs or are affected by the AI system, and, as a result, this category includes but goes beyond users. For example, an AI system used for predictive maintenance has both users (maintenance personnel) and stakeholders (logisticians, pilots, planners, etc.). Identifying stakeholders is especially important in contexts where AI is used to make predictions about people. It is best practice to get input from people who are going to be evaluated by an AI system to ensure that they understand and are comfortable with its intended use.

Some national security contexts may prevent stakeholder consultation. In these cases, mission users should identify an appropriate proxy who represents the needs, perspectives, and interests of the individuals affected by the AI system. In cases where this is not possible, end users and mission owners should comprehensively document the specific conditions, circumstances, and prohibitions that prevent the consultation of any given stakeholder. End users and development teams should also seek well-documented approvals and authorizations from the mission owners or decision makers to ensure that their actions are validated by appropriate authorities and documented accordingly.

[13] Google, "User Needs + Defining Success," in People + AI Guidebook, https://pair.withgoogle.com/chapter/user-needs/

RESPONSIBLE MISSION OWNER

AI systems cannot be responsible for outcomes—humans must always bear responsibility. In particular, irreversible decisions that affect a person's life, quality of life, health, or reputation should be made by a human, not a machine.

The mission owner is in charge of defining success and is accountable for ensuring that the capability meets operational, organizational, and ethical requirements. This person should sit within the project execution team and have appropriate understanding of both the technical and operational aspects of the project. This is the person responsible for navigating trade-offs and ensuring clear communication of objectives both internally and externally at each phase of the project.

The mission owner should also operate and make decisions in consultation with legal counsel, who can adjudicate legal concerns that arise at the earliest stage of the planning process. Legal input provides guard rails that ensure an AI system is designed and developed in compliance to all relevant laws and regulations. Legal consultation should occur throughout all three phases of the AI technology development cycle.

4. Have you conducted harms modeling to assess likelihood and magnitude of harm?

This includes, but is not limited to: injury, denial of consequential services, infringement on human rights, erosion of social and democratic structures, and consideration of particular groups which may be advantaged or disadvantaged in the context in which you are deploying the capability.

Harms Modeling, as defined by Microsoft's Ethics & Society team, "is a practice designed to help you anticipate the potential for harm, identify gaps in product that could put people at risk, and ultimately create approaches that proactively address harm."[14]

The first step is to identify a list of potential harms. Harms may be organized into the following categories:

- Physical injury: how the capability can injure persons, e.g. misdiagnosis of an illness, failure of critical components, incorrect targeting, etc.
- Psychological injury: how the capability can cause severe psychological distress, e.g. intrusive monitoring, identity theft through deepfakes, and misattributions from failed facial recognition
- Opportunity: how the capability could limit access to important resources, services and opportunities, e.g. biased hiring or promotion algorithms, discriminatory benefit allocations, and negative impacts on groups without digital access
- Human rights and civil liberties: how the capability could impact human rights and civil liberties, e.g. violation of privacy, loss of due process, limitation of free choice, disparate impact, and disparate treatment.
- Environmental impact: how the capability can produce harmful environmental effects, e.g. unnecessarily complex algorithms creating high energy demands
- Social and democratic values: how the capability can erode or violate social and democratic values, e.g. manipulation through disinformation or behavior exploitation, stereotyping, etc.

There are other approaches that can be used to identify broad potential harms to people using the system, people interacting with data in the system, people whose information is managed by the system, and people who may be unintentionally harmed due to system operation.[15] This is a significant effort and the entire development team should work through these activities and be speculative and imaginative in identifying both beneficial and harmful outcomes from these systems.

AI systems often have far-reaching impacts both when they function well, and when they function poorly. For example, a well-functioning system that automates tasks previously performed by a human could have a positive impact if it allows that person to

[14] Microsoft. "Foundations of Assessing Harms." https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

[15] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. "Guidelines for Human-AI Interaction." In *CHI Conference on Human Factors in Computing Systems Proceedings* (CHI 2019), Glasgow, Scotland Uk. May 4–9, 2019,ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300233; Carol J. Smith. 2019. "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development." arXiv:1910.03515 Retrieved from https://arxiv.org/abs/1910.03515; Dan Brown. 2018. "UX in the Age of Abusability". *Green Onions* (Blog). September 18, 2018. Retrieved September 13, 2019 from: https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13; Michael Chapman, Ovetta Sampson, Jess Freaner, Mike Stringer, Justin Massa, and Jane Fulton Suri. 2018. "Data, Ethics, and AI: Practical activities for data scientists and other designers." *Medium* (Blog). October 12, 2018. Retrieved September 13, 2019 from: https://medium.com/ideo-stories/data-ethics-and-ai-276723a1a2fc; Casey Fiesler. 2018. "Black Mirror, Light Mirror: Teaching Technology Ethics Through Speculation." *NEXT* (Blog). October 15, 2018. Retrieved September 13, 2019 from: https://howwegettonext.com/the-black-mirror-writers-room-teaching-technology-ethics-through-speculation-f1a9e2deccf4

focus on other tasks, or a negative impact if it renders that person redundant. A predictive maintenance system that performs poorly could impact not only maintenance personnel but also pilots, logisticians, planners, and others.

It is particularly important to pay attention to the distribution of advantages and disadvantages when AI systems are used to make predictions about people. For example, an algorithm trained on past promotion decisions will inherit any explicit or implicit historical biases, and could disadvantage underrepresented groups going forward. Remember that when accuracy is used as the metric for success, one is considering how a model performs overall rather than on individual categories. Suppose one wants an algorithm that identifies both cats and dogs. If there are 80 cats and 20 dogs in a population, that algorithm would have the same accuracy whether it correctly identifies 80 cats and no dogs or 60 cats and 20 dogs. It is important to note that system performance can, and often is, much worse on groups that are underrepresented in the training data unless intentional steps are taken. When such “imbalance” in datasets occurs—and it often does in practice!—it is crucial to design test, evaluation, and monitoring procedures to ensure that the algorithm both initially and continually accounts for such imbalance appropriately.

Some other common types of undesirable bias include, but are not limited to:

- Sample bias: This occurs when data is not representative of real world conditions.
- Automation bias: This occurs when human operators put too much stock in algorithmically generated outputs vs. human judgments.
- Label bias: The choice of how data is labeled may have a deleterious impact on certain individuals, e.g, dividing a population into Black/white or man/woman would discriminate against individuals who do not self-identify with either category.
- Prejudice by proxy: Sensitive attributes like race and gender may be highly correlated with other attributes (e.g. zip code). Consequently, simply removing information about sensitive attributes does not guarantee protection against unwanted bias.

Potential questions for program managers/vendors:

- Conducted a harms analysis assessing risk of:
  - physical harm
  - psychological harm
  - opportunity loss
  - human rights and civil liberties violations
  - environmental impact
  - erosion of social and democratic values
- For each harm identified, consider:
  - severity (how big of an impact?)
  - scale (how wide of an impact?)
  - probability (how likely is the harm to occur?)
  - frequency (how often could the harm occur?)
- Consider the potential damaging effect of uncertainty or errors to different groups:
  - What are realistic worst-case scenarios in terms of how errors might impact society, individuals, and stakeholders? Ideally, his should be addressed for each stakeholder.
- What are the operational risks if things go well vs. if they go wrong?
  - If things go well: What would those impacts look like at the individual and community levels?
  - If things go wrong: What are those impacts at the individual and community levels? How might these individuals/ communities be prevented from accessing services?

5. Have you identified the process for system rollback and error identification/correction ?

SYSTEM ROLLBACK

An AI capability that works at deployment may fail later for a variety of reasons (e.g. model drift). It is critical that the project have, from the outset, a plan for what to do when and if this occurs. If a task previously performed by a person is now fully automated, and that automation fails, it may be necessary to revert to a manual process, and fail safely first. Thus, it is critical to maintain personnel who can perform the task in the absence of an AI capability for a substantial period after deployment.

A robust set of automated and verified tests should be planned, along with a schedule to run those tests. Drift in the outcomes of these tests provides a set of alarms that inform whether the AI capability is performing sub-optimally and can also help with diagnosing exactly what is going wrong.

Potential questions for program managers/vendors:

- Who in the DoD (name and/or specific role) will be able to monitor the system and how?
- Who in the DoD will be able to control and deactivate the system (if necessary) and where will that process be documented?
- How will the change be communicated to end users and other stakeholders?

ERROR IDENTIFICATION / CORRECTION

A major differentiating factor between AI capabilities and traditional software is that it is generally not possible to test all possible paths the AI capability could encounter. As a result, it is critical not only to establish what types of errors would be important in an operational context, but also to define processes for detecting them (e.g. human audit, dual phenomenology, etc.) and for correcting them. Both detection and correction can involve algorithmic and human or organizational process components, but a clear list of error modes and potential remedies should be created during the planning phase. If any of the error modes is both particularly concerning and difficult to mitigate, development of the AI capability should be reconsidered, or the task and/or deployment environment should be reframed.

Potential questions for program managers/vendors:

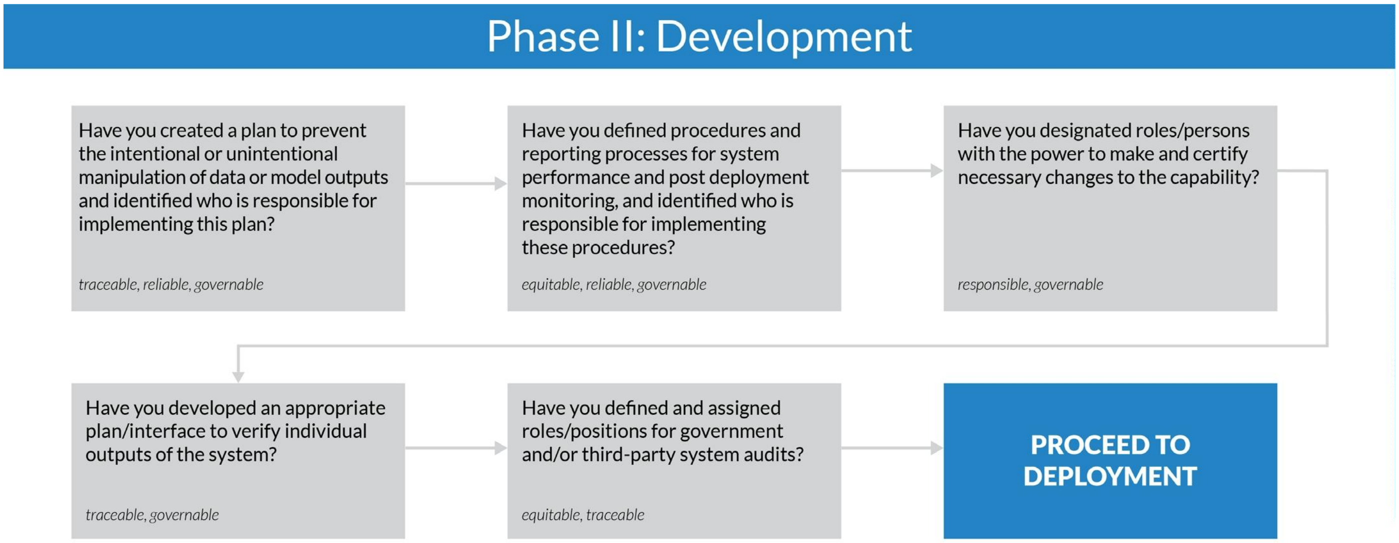
- Have you determined the process for error identification and correction?
- Have you created a list of potential error modes and remedies?
- Who will have the power to decide on necessary changes to the capability during the design phase, pre-launch, and post-launch?
- Who and how will those changes be certified?

Phase 2: Development

The development worksheet is to be completed and updated jointly by the commercial vendor team and the government, with DIU support. The planning worksheet should already be complete and may be updated as needed.

Directions: Respond to the following questions in the order they are presented and include notes about your conversation(s) with regard to applicability for development efforts. Include descriptions of what has already been completed and what work is left to be done (if applicable). Include context as appropriate, such as a timeline for completion and current status. Please provide a justification if the question is not applicable to the project or the issues raised will be resolved at a later date.

DEVELOPMENT WORKSHEET // DEVELOPMENT PROCESS FLOW



Phase II: Development

Have you created a plan to prevent the intentional or unintentional manipulation of data or model outputs and identified who is responsible for implementing this plan?

traceable, reliable, governable

Have you defined procedures and reporting processes for system performance and post deployment monitoring, and identified who is responsible for implementing these procedures?

equitable, reliable, governable

Have you designated roles/persons with the power to make and certify necessary changes to the capability?

responsible, governable

Have you developed an appropriate plan/interface to verify individual outputs of the system?

traceable, governable

Have you defined and assigned roles/positions for government and/or third-party system audits?

equitable, traceable

PROCEED TO DEPLOYMENT

QUESTIONS

1. Have you created a plan to prevent the intentional or unintentional manipulation of data or model outputs and identified who is responsible for implementing this plan? Lay out your plan.
2. Have you defined procedures and reporting processes for system performance and post deployment monitoring, and identified who is responsible for implementing these procedures? Define these standard operating procedures.
3. Have you designated roles/persons with the power to make and certify necessary changes to the capability? Identify these individuals.
4. Have you developed an appropriate plan/interface to verify individual outputs of the system? Explain your plan.
5. Have you defined and assigned roles/positions for government and/or third-party system audits? Explain your approach.

COMMENTARY ON DEVELOPMENT WORKSHEET

The following is intended to accompany the questions for the development flow and provides additional context for the questions asked in order to guide the team through the evaluation process.

1. Have you created a plan to prevent the intentional or unintentional manipulation of data or model outputs and identified who is responsible for implementing this plan?

VERIFY THAT ADVERSARIES CANNOT GAIN ROOT OR QUERY ACCESS TO THE MODEL

Root access refers to the ability to not only acquire but change the characteristics of a dataset or model. If an adversary has root access they may be able to perform data poisoning—intentionally distorting data such that models trained on that data fail in operation (often in specific ways).[16] Data poisoning at the root level is difficult (if not impossible) to identify, and can have disastrous operational consequences. For example, an algorithm trained on maliciously altered network traffic data may be unable to detect an adversary's cyber attacks even though model performance is, from the user's point of view, incredibly high.

Query access refers to the ability to input and receive outputs from a model. If an adversary has query access, they may be able to infer properties of the model which could then become the basis for intentional manipulation. Adversarial attacks on opaque AI systems occur when the input to a model is manipulated to either produce an erroneous (untargeted attacks) or specific (targeted attacks) output. For example, researchers have found that manipulating a small number of pixels on image inputs can cause an otherwise highly performing algorithm to catastrophically fail. In another example, researchers placed small amounts of tape to successfully manipulate a state-of-the-art computer vision model into misclassifying a stop sign as a 60 mph speed sign.

There are various tools and techniques to counter data poisoning and adversarial attacks on both opaque and transparent systems. It is critical that such precautions are implemented in instances where an adversary could gain (root) data, query, or model access.

Potential questions for program managers/vendors:

- Who currently has or will have root access to the dataset and model?
- How are permissions for root access managed?
- What do you see as the most likely adversarial attacks on your system?
- What type of attack would be catastrophic for your system? What are the standard operating procedures for when this happens? Who do we report to if it does? Is there a threshold to meet?

2. Have you defined procedures and reporting processes for system performance and post deployment monitoring, and identified who is responsible for implementing these procedures?

SYSTEM PERFORMANCE

As operational requirements and context evolve, models must evolve as well. A model trained on a dataset with particular characteristics may fail because those characteristics are no longer representative of reality. For example, a model trained to identify

[16] Ali Shafahi, W.Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, Tom Goldstein. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," in *32nd Conference on Neural Information Processing Systems, NuerIPS, Montreal, Canada*, 2018,1-11 https://proceedings.neurips.cc/paper/2018/file/22722a343513ed45f14905eb07621686-Paper.pdf

rocket sites from electro-optical imagery may fail either because site architecture is now different or the sensor used to collect imagery has changed. Alternatively, a model trained to perform a certain task may fail because the task it is now expected to perform is different. For example, a model trained to classify cats and dogs may continue to perform well when only cats and dogs are present, but fail operationally because llamas have also entered the scene.

Edge cases present a particular challenge because it is highly unlikely that they will be exhaustively identified and accounted for prior to model deployment. Consequently, it is critical to continuously and quantitatively test system performance using metrics that are appropriate to the current operational task. If either that task or the context in which that task is performed has changed, it is likely that the model requires changing too.

Potential questions for program managers/vendors:

- What are the primary characteristics of your training dataset?
    - In what ways is it applicable to your deployment context? Where is that recorded?
    - In what ways is it not applicable to your deployment context? Where is that recorded?
    - How was your labeling schema defined, and is it likely to change?
- What data do you wish you had that would increase the quality of your training dataset?
- What procedures exist to improve data quality over time?
- Who on your team is tracking changes to deployment context over time?
- What is the process for deciding when to retrain a model and who is responsible for that decision?

POST DEPLOYMENT MONITORING

Post deployment monitoring includes upversioning and downversioning capabilities. Upversioning refers to the replacement of the current capability with a newly developed version of the capability. This may be desirable if, for instance, the type of data being classified by a machine learning model changes over time, and a new model version has been trained to account for this issue. Downversioning refers to the replacement of the current capability with a previous version of that same capability. This is incredibly important, as if one finds an important error mode in a currently deployed version of a capability, one should be able to quickly revert to the most recent “stable” version.

Post deployment monitoring is required for AI capabilities while in deployment. It is insufficient to test a model once and assume it will continue to function when deployed. Thus, a plan for how models will be tested after they are deployed—what tests will be run, and the critical performance thresholds below which the model must not fall in order to remain deployed—should be defined before the model is deployed. In traditional software, these issues are often handled via continuous integration/continuous deployment processes.

The person who and/or a specific role that is responsible for post-deployment monitoring must be identified.

Potential questions for program managers/vendors:

- What is the cadence for testing after models are deployed?
- Who is deciding on metrics for success?
- What is the requirement to maintain or keep previous versions?
- What types of situations will drive your team to downversion? Who makes that decision?
- How will that impact the end users and stakeholders? How will they be notified?
- What is the process for reacting when error modes are discovered? Who is involved in addressing errors?

REPORTING AND ADDRESSING UNDESIRABLE SYSTEM BEHAVIOR

Even the most state-of-the-art AI system will make mistakes. This does not mean that the system is not operationally useful—indeed, AI systems should only ever be employed in cases where there is tolerance for a certain degree of error or where appropriate redundancies and safeguards are in place.
AI systems tend to be highly complex, which can result in unintended or undesirable behavior. The cause of such behavior may not always be obvious, and so it is critical that care is taken to document when and under what circumstances this occurs. Data collected about system behavior can be used to both improve system performance and more specifically tailor the performance envelope of the AI system (e.g. conditions in which the AI system can reliably perform). This data collection and system characterization should also continue during capability deployment.

Have error modes for each task been identified? AI systems can have a wide range of error modes. Computer vision systems for classifying images can, for instance, consistently confuse two particular classes or misclassify all images of a particular type. The implications of each possible error mode in the context of the operational workflow must be identified in order to assess how tests and usage protocols for the capability should be designed.

What is the plan to mitigate the impact of each error mode on operational outcomes? Once error modes are identified, developers should clearly state how each will be addressed. This could occur through a combination of algorithmic, human, and workflow considerations, but a clear assessment of the cost of each type of error and the plan for mitigation of each type of error must be performed and continuously updated during capability development.

Potential questions for program managers/vendors:

- What error modes surfaced during testing with subject-matter experts and/or end users?
  - How did your team learn from them? What changed in the process by which the task is done as a result?
  - How did the system learn from them? What changed in the system as a result?
- What did the process for identifying errors across all system tasks entail?
- What error modes have the largest impact on your users? What error mode is likely to occur the most frequently?
- Who is tracking error modes over time? Where is that information stored?
- How will your team debrief errors?

3. Have you designated roles/persons with the power to make and certify necessary changes to the capability?

Just as mechanical systems require regular inspections, AI systems should be subject to periodic review and re-certification by appropriately trained and accountable personnel. The proper individuals for such roles will depend upon the context of use and the nature of the system; in some cases it may be appropriate to delegate responsibility to the program officer whereas, in others, it may be necessary to bring in staff with deeper technical expertise. In any event, accountability should be traceable to a single individual who either possesses or has access to the required expertise to assess—and is empowered to make any necessary changes to—current performance of the AI system.

Potential questions for program managers/vendors:

- Is there a specific person (or role) designated to track, monitor, and certify changes to the system while in development?
  - Does that person (or role) have the requisite authority to assess changes, and, if necessary, authorize and executive corrective actions when needed?
  - Does that person (or role) have full visibility (administrator privileges) on the system inputs, outputs, and evaluation metrics used to track and monitor the system during development?
  - Has that person (or role) developed procedures that ensure system continuity if they are replaced?

4. Have you developed an appropriate plan/interface to verify individual outputs of the system?

Many AI systems are intended to support decisions that can have extremely high cost if made incorrectly. As a result, not only should a system be tested on aggregate performance—i.e. how well the system performs on average—but processes should also be put into place to ensure the validity of any individual outputs used within an operational workflow.

In some cases, such processes are not required, though may be desirable for traceability and transparency. For instance, if a model is simply prioritizing the order in which an analyst will review satellite images—and the analyst will, in fact, review each of those images—it is unlikely that each individual prioritization prediction should be reviewed.

However, if a model is making a prediction about whether or not a potential target has been detected on a satellite image, it would be appropriate to ensure that an analyst reviews each positive target prediction manually, or that dual phenomenology is used to confirm the output.

Potential questions for program managers/vendors:

- How will the system enable the verification of individual outputs?
- How will decisions be made about traceability and transparency with regard to outputs?
- Who will make those decisions?
- Which outputs will be verifiable by end users? Which outputs will only be accessible by administrators?

5. Have you defined and assigned roles/positions for government and/or third-party system audits?

If the vendor facilitates third-party auditing, the government should clearly establish the goals and procedures associated with the audit, and define a verifiable reporting structure that can be used by the third-party auditor to confirm that items in the development workflow have been appropriately addressed. The third-party auditor should be able to conduct the audit without opening the system to prevent unwarranted manipulation.

Models can be audited in multiple ways, ranging from internal code and training process reviews to fuzzing and deterministic testing, and different applications will require different degrees of capability auditing.

Will the vendor allow the government to audit directly? If the vendor allows the government to audit the capability directly during development, government representatives should define a clear auditing plan for evaluating how each of the previous questions in this flow chart has been answered.
It is a red flag if the vendor refuses to allow third-party or government system audits without a very compelling reason.

Potential questions for program managers/vendors:

- What method will be used to enable auditing of the system by a third party or the government?
- What are the goals and procedures for audits?
- What will and will not be audited?
- How will audits be reported (format, timeline, etc.)?

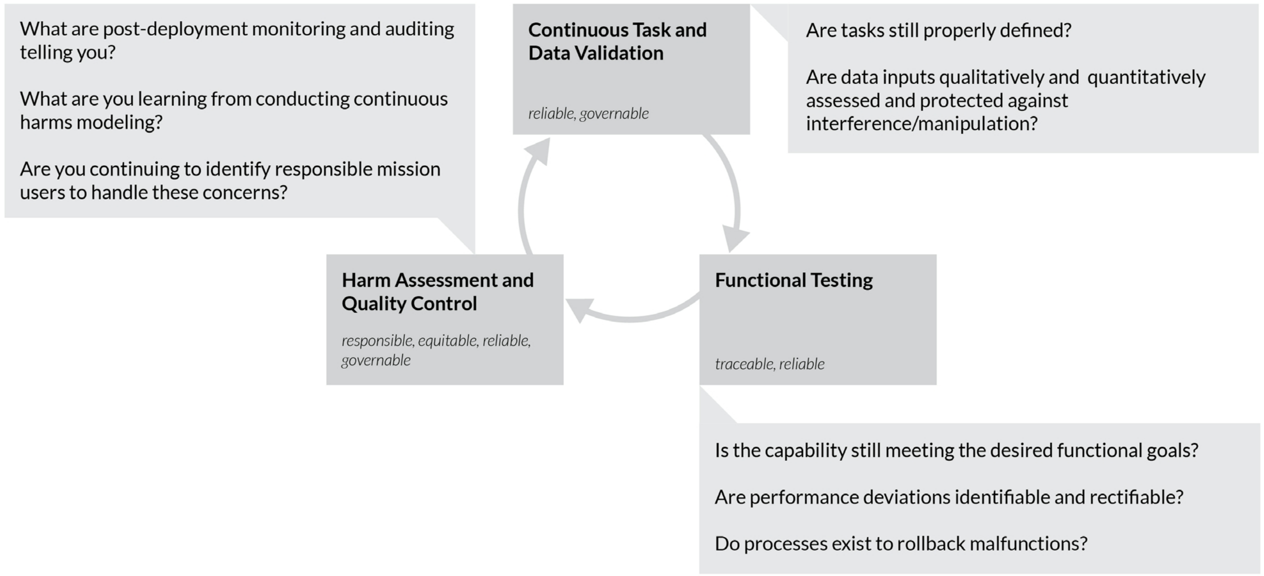
Phase 3: Deployment

The deployment worksheet is to be completed and updated jointly by the commercial vendor team and the government, with DIU support. The planning worksheet and the development worksheet should already be complete and may be updated as needed.

Directions: These questions build on the work done in both the planning and development worksheets. Respond to the following questions in the order they are presented and include notes about your conversation(s) with regard to applicability for deployment efforts. Include descriptions of the context, progress, and overall status as appropriate. If a question or topic is not applicable to this project, please include a justification from previous worksheets.

DEPLOYMENT WORKSHEET // DEPLOYMENT PROCESS FLOW

Phase III: Deployment



Continuous Task and Data Validation

reliable, governable

Are tasks still properly defined?

Are data inputs qualitatively and quantitatively assessed and protected against interference/manipulation?

Functional Testing

traceable, reliable

Is the capability still meeting the desired functional goals?

Are performance deviations identifiable and rectifiable?

Do processes exist to rollback malfunctions?

Harm Assessment and Quality Control

responsible, equitable, reliable, governable

What are post-deployment monitoring and auditing telling you?

What are you learning from conducting continuous harms modeling?

Are you continuing to identify responsible mission users to handle these concerns?

QUESTIONS

Continuous Task and Data Validation

1. Are tasks still properly defined?
2. Are data inputs qualitatively and quantitatively assessed and protected against interference/manipulation?
3. Functional Testing
4. Is the capability still meeting the desired functional goals?
5. Are performance deviations identifiable and rectifiable?
6. Do processes exist to rollback malfunctions?

Harms Assessment and Quality Control

7. What are you learning from conducting continuous harms testing?
8. What is post-deployment monitoring and auditing telling you?
9. Are you continuing to identify responsible mission users to handle these concerns?

COMMENTARY ON DEPLOYMENT WORKSHEET

The following is intended to accompany the questions for the deployment flow and provides additional context for the questions to guide the team through the evaluation process.

Continuous task and data validation

1. Are tasks still properly defined?

If operational demands change, AI systems need to change too. A model that was trained to distinguish between dogs and cats should no longer be used if the task is now to distinguish between dogs, cats, and llamas. Properly documenting classification schema, training data, optimization metrics, etc. is absolutely essential and should be compared to current operational requirements to ensure that there is an appropriate match.

Potential questions for program managers/vendors?

- How is the capability evaluated to ensure that it still delivers desired outputs?

- How are changes to operational requirements tracked to ensure that the system continues to deliver desired outputs?
- How are changes to data inputs or outputs, evaluated to ensure that the system delivers optimal results

Continuous task and data validation

2. Are data inputs qualitatively and quantitatively assessed and protected against interference/ manipulation?

The quality and origin of data used during model development may be (and often is) different from data in a deployed context. Recording these changes is required insofar as they may require a re-examination of data preparation procedures (e.g. extract-transform-load, normalization, or cleansing). A paper trail will ensure that future users are able to identify when and how deviations in data provenance and quality occur, which could be required for corrective action (e.g. model rollback).

Potential questions for program managers / vendors:

- How will new data for the system be assessed and managed?
- How will adjustments in data preparation be recorded?
- Who will manage this work and who will have access to it?

Functional testing

3. Is the capability still meeting the desired functional goals?

Machine learning is a rapidly advancing technology. Model performance should be continually monitored and compared to both state-of-the-art and operational requirements. The lifetime of models should be measured in weeks or months, not years.

Model performance on its own may not be a reliable indicator of whether the capability is still providing value. The model may no longer be relevant to the current requirements, or other aspects of the system (e.g. user interface) may inhibit the full realization of system benefits. Consequently, a periodic review should be conducted to consider the quantitative measurements of the model and to assess how well the capability functions as a whole.

Potential questions for program managers/vendors?

- How are models examined to ensure they consistently address the desired functions?
- How are changes to model performance recorded and tracked?
- Who will manage periodic reviews of the capability and its performance?

Functional testing

4. Are performance deviations identifiable and rectifiable?

Performance degradations should be defined by the metrics used during deployment; if these metrics need to be updated, it implies that the model may not be well suited for the task and rollback should be considered.

Potential questions for program managers/vendors?

- How are performance changes tracked and assessed to identify deviations that impact the output?
- If performance deviations adversely affect the model output, is a corrective process in place to rectify these changes?
- Who will manage the correction process if deemed necessary?

Functional testing

5. Do processes exist to rollback malfunctions?

If the system is not performing as expected or is not functional, rollback should be considered, and post-deployment monitoring should be increased. If the plan for rollback is not functional, resources should be immediately allocated to address the issue, as operational success would no longer be achievable without the AI capability.

Potential questions for program managers/vendors:

- Has there been a need for a rollback? If yes, what improvements to the plan would support the team?
- If not, what are the concerns (if any) with the existing rollback plan?
- Who manages or is responsible for system rollbacks and do they have the requisite access to the system?

Harms assessment and quality control

6. What are you learning from conducting continuous harms testing?

Models can perform as desired, but could still have unintended effects on the overall workflow. For instance, if a model is doing well at automatically categorizing images, analysts assigned to review may become less engaged, leading to an increased overall error rate from the classification workflow. Similarly, occasional model errors can result in user distrust (particularly if they are unexplained in existing documentation).

Models can slowly change performance characteristics over time. Consistent evaluation for disparate impact and treatment is critical to ensuring that such problems do not occur.

Potential questions for program managers/vendors:

- How are performance outputs evaluated to identify potential harms during deployment?
- Are harms assessments conducted on a regular, recognizable and predictable schedule?
- Who is responsible for managing harms testing and evaluation?

Harms assessment and quality control

7. What is post-development monitoring and auditing telling you?

AI systems can fail for a multitude of reasons that make continuous and quantitative monitoring of system performance critical. As discussed in the development phase, it is critical that all AI systems have a plan for continually monitoring performance, and recording and responding to undesired system performance.

The goal of post-deployment monitoring is to ensure that the capability functions as designed. Because AI systems are difficult to exhaustively test, one must ensure that consistent post-deployment evaluation is performed to identify potential errors before they become problematic, mitigate the potential impact of those errors, and provide clear guidance as to how models should be updated before redeployment. Additional types of tests should be considered as time goes on or as additional potentially undesirable behaviors are identified.

Potential questions for program managers / vendors:

- How have monitoring and auditing systems supported your work?
- How might they be improved to support future use?
- What are you finding to be the most common issues?
- When issues have been identified, what has been difficult to manage?
- What additional testing is needed to meet the needs?

DEFENSE INNOVATION UNIT

ACCELERATING COMMERCIAL TECHNOLOGY FOR NATIONAL SECURITY

www.diu.mil