DEFENSE INNOVATION UNIT

# Phase III: Deployment Worksheet for DIU AI Guidelines

**Project Name**: _____

**DIU Project Manager**: _____

**DoD Point of Contact**: _____

**Updated on**: _____

## Contents



## AI Guidelines Process Overview

The Defense Innovation Unit has created the AI Guidelines and process to help guide thinking and surface potential issues sooner, rather than later to avoid unintended consequences in creating AI systems. The process includes completing worksheets for planning, development and deployment efforts. These are not a legally binding documents nor are they intended to supplant or replace existing laws and regulations.

This work is based on the five DoD Ethics Principles for the development and use of artificial intelligence that were adopted by the Secretary of Defense in 2020:

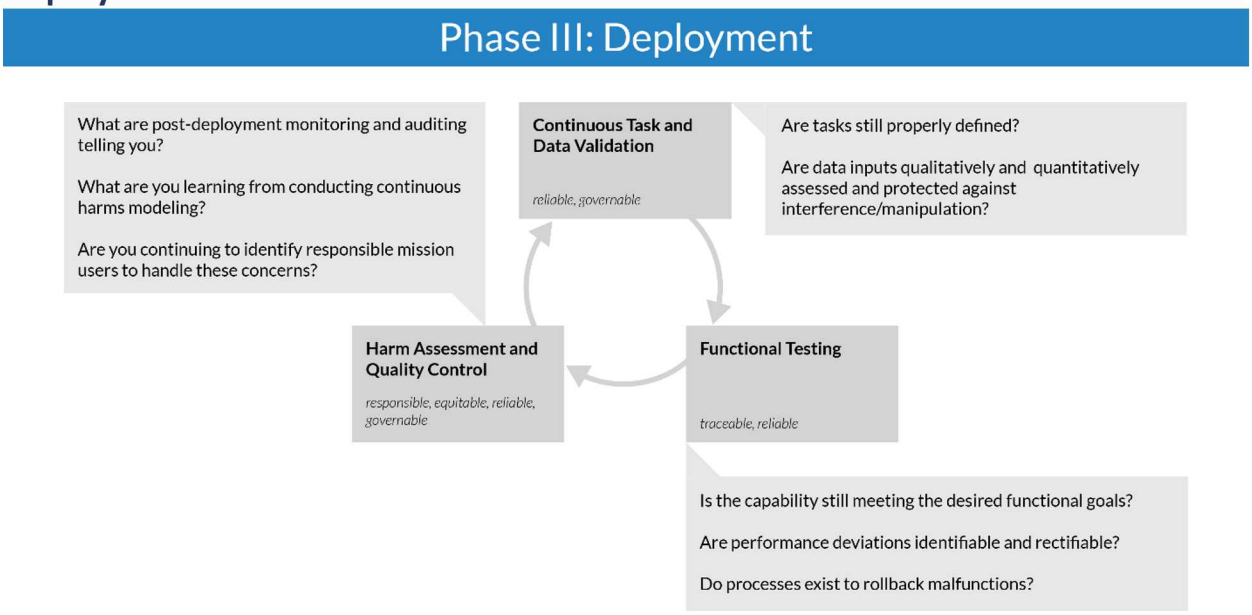
> ***Responsible.*** *DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.*
>
> ***Equitable****. The Department will take deliberate steps to minimize unintended bias in AI capabilities.*
>
> ***Traceable****. 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.*

The deployment worksheet (this document) 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 question 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 the topic is not applicable to this project, please include a justification from previous worksheets.

# Deployment Worksheet

## Deployment Process Flow




## Continuous Task and Data Validation

### 1. Are tasks still properly defined?

Commentary

---

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

Commentary

---

## Function Testing

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

Commentary

---

### 4. Are performance deviations identifiable and rectifiable?

Commentary

---

### 5. Do processes exist to rollback malfunctions?

Commentary

---

## Harms Assessment and Quality Control

### 6. What are post-deployment monitoring and auditing telling you?

Commentary

---

### 7. What are you learning from conducting continuous harms modeling?

Commentary

---

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

No commentary provided

---

## 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, outputs, and/or end-users, evaluated to ensure that the system delivers optimal result?

*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 that not only considers the quantitative measurements of the model, but also assesses 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 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, 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 for how models should be updated before redeployment. Additional types of tests should be considered as time goes on 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?

*Harms Assessment and Quality Control*

## 7. What are you learning from conducting continuous harms modeling?

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?