DEFENSE INNOVATION UNIT

Phase I: Planning 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:

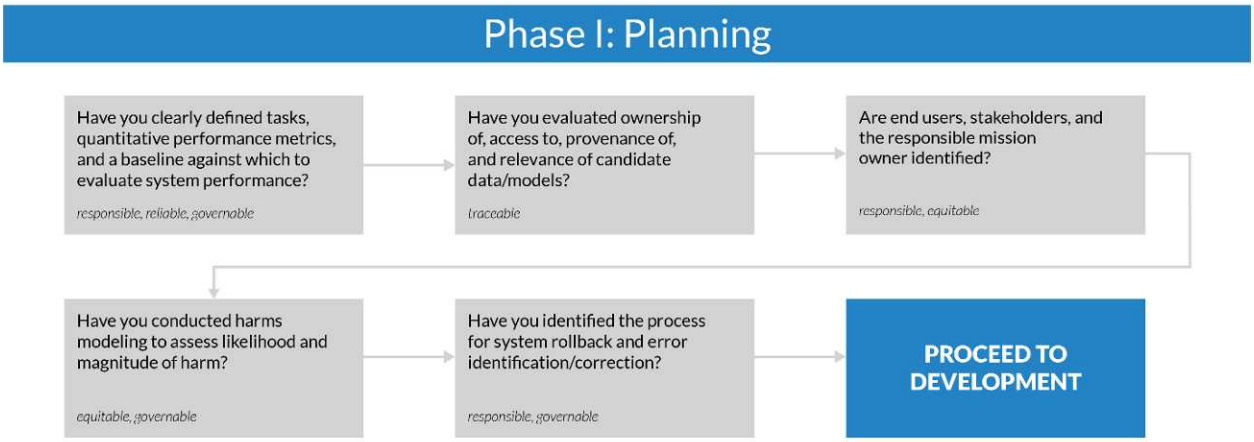
> ***Responsible.*** *DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.*
>
> ***Equitable****. The Department will take deliberate steps to minimize unintended bias in AI capabilities.*
>
> ***Traceable****.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

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



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

Commentary

---

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

Commentary

---

3. Are end users, stakeholders and the responsible mission owner identified?
Identify these groups or individuals.
Commentary

---

4. Have you conducted harms modeling to assess likelihood and magnitude of harm?
Explain your approach. What did you find?
Commentary

---

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

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 the team 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[1]. 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[2]

- 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 fraudulent3, 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).

[1] 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

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

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

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, the inputs and outputs are not identified. One must always be wary of AI projects that seek to "improve" performance -- 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.

Potential questions for program managers / vendors:

- Is AI suitable to 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 you cannot state your objective mathematically, you 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, a 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 you know whether your 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[4], 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 tradeoff between accuracy and speed. Knowing the baseline will allow the team to navigate these tradeoffs by understanding what is currently acceptable 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 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, and it 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.

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.)?

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

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 where the data was sourced from (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 is 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

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?

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 determine 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[5]. 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.

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

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 using the system outputs and/or are affected by the AI system, and so includes but goes beyond users. For example, an AI system used for predictive maintenance has both users (maintenance personnel) and stakeholders (logistics, 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.

Responsible Mission Owner

AI systems cannot be responsible for outcomes – humans must always bear responsibility. In particular, final (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 stage of the project.

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."[6]

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

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, misattributions from failed facial recognition
- Opportunity: how the capability could limit access to important resources, services and opportunities, e.g. biased hiring / promotion algorithms, discriminatory benefit allocations, negative impacts on groups without digital access
- Human rights & 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, stereotype

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.[7] 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 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, logistics, 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 less well represented 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

---

[7] 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

20 dogs. It is important to note that system performance can, and often is, much worse on groups that are less well represented 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 / errors to different groups:
    - What are realistic worst-case scenarios in terms of how errors might impact society, individuals, and stakeholders? This should ideally 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 later fail 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 thus 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 if 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/organizational process components, but a clear list of error modes and potential remedies should be created in the planning stage. 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 stage, pre-launch, and post-launch?
- Who and how will those changes be certified?