MITRE

Risk Discovery Protocol for AI Assurance (v1.0)

Guidance for Administering the Protocol

Paul Ward
Jeff Stanley
Ron Ferguson
Colin M. Gladding
Kevin J. Burns

October 2024

Approved for Public Release; Distribution Unlimited. Public Release Case Number 24-2963.

McLean, VA

Contents

1 Purpose of this Document

The purpose of this document is twofold: To provide an overview of MITRE's AI Assurance and Discovery (AIAD) Lab's *Risk Discovery Protocol for AI Assurance (RDP-AIA)*; and to offer guidance on how to administer this protocol. The intention is that any responsible party given minimal context could administer the protocol using this document as a guide.

2 Introduction: The Risk Discovery Protocol for AI Assurance

2.1 Terminology

In this document, the team that administers the RDP-AIA will be referred to as the AIAD Lab team or just the **AIAD Lab team** (capital *L*). The Stakeholder who comes to the AIAD Lab with assurance concerns and who is the direct customer of the RDP-AIA will be referred to as the **Stakeholder** (capital *S*).

2.2 Overview

The RDP-AIA is a standardized methodology explicitly designed to support a Stakeholder's ability to: (a) decompose a specific mission problem and the associated AI solution that addresses that challenge, (b) aid identification and discovery of known and potential AI assurance (AIA) needs relevant to the proposed solution or Sponsor challenge, and (c) provide a means to prioritize identified AIA needs based on the level of risk posed.

The RDP-AIA was developed for use in MITRE's AIAD Lab and represents the first step in the AIAD Lab's core function: To discover/prioritize AIA concerns prior to measuring and mitigating the associated risks. As such, this protocol constitutes the "front-end" of AIAD Lab services, and acts as the first means of interaction with the AIAD Lab for Stakeholders, both internal and external to MITRE.

The RDP-AIA is comprised of four phases, with the following objectives:

- **UNDERSTAND:** Unpack the mission problem, proposed AI solution, and expected effects.
- **IDENTIFY:** Elicit Sponsor-specific AIA needs and potential impacts across the lifecycle.
- **DISCOVER:** Explore the AIA landscape, similar use cases, broader assurance issues, and possible effects.
- **PRIORITIZE:** Assess current and future risks and risk tolerances and prioritize AIA needs.

The RDP-AIA is implemented as a four-phase, Stakeholder interview and data-collection process, with each phase providing outputs that serve as inputs to subsequent phases, culminating in an RDP-AIA Priority Report (henceforth, Priority Report). This four-phase protocol is represented by the flowchart in Appendix A. It can be administered in person or remotely.

The focus of this guidance document is on supporting use of the RDP-AIA in collaboration with a Stakeholder. In particular, the focus is on describing and providing guidance to administer the interview questions asked and the associated activities employed during each phase of the protocol. These phases together with the corresponding steps, interview questions, activities, and associated guidance are described in more detail in the following sections.

This protocol document also contains a description of several key artifacts that are integral to the RDP-AIA. A brief description of each artifact is provided below; each will be introduced in their respective phase of the protocol:

- **Use Case Description**: A comprehensive yet simplified description of each use case generated from data gathered during pre-interview and during the UNDERSTAND phase of the RDP-AIA. A template for capturing this information is shown in Appendix A.
- **AIA Landscape:** An illustrative representation of the core AIA needs categories and specific needs used to guide AIA needs exploration in the DISCOVER phase (see Appendix A).
- **Glossary:** Definitions of AIA needs categories and specific needs contained in the AIA Landscape. The Glossary serves as a necessary reference point to minimize talking at cross purposes throughout the RDP-AIA, aid exploration of AIA landscape in the DISCOVER phase, and facilitate assessment of associated risks in the PRIORITIZE phase (see Appendix E).
- **Risk Assessment Tools:** Rating scales used to assess severity and likelihood of impacts that might result from not assuring a particular AIA need, and a risk matrix to assess overall risk in the PRIORITIZE phase (see Appendix H).
- **RDP-AIA Priority Report:** A summary of the findings from administration of the RDP-AIA with a particular Stakeholder, highlighting the AIA priorities and risks for the use case in question and associated recommendations. A template for capturing relevant information in the Priority Report, following the PRIORITIZE phase and completion of the interview, is shown in Appendix I. On completion of the protocol, the Priority Report will be appended to the respective Use Case Description.
- **Use Case Repository[1]:** A catalogue of completed Use Case Descriptions (i.e., from Stakeholders that have completed the RDP-AIA), appended with the respective Priority Report, and indexed by key features of the respective Sponsor challenge/AI solution. Entries in the Use Case Repository with attributes similar to the current use case can be used to help guide exploration of potential AIA concerns during the DISCOVER phase. (See Appendix C for details.)

Each of the above artifacts are used in one or more of the phases of the RDP-AIA, and each one serves as input to and/or output for a subsequent phase (as described below).

The RDP-AIA Priority Report constitutes the primary output from the RDP-AIA and serves three purposes: It provides a summary of the AIA priorities and recommendations to the interviewed

[1] The Use Case Repository is a future effort that will emerge as a natural consequence of multiple Stakeholders engaging in an RDP-AIA.

Stakeholder, input to the *Use Case Repository* to enrich future projects that have problem or solution features in common with the past use cases, and input to the next step of the AIAD Lab process: Measuring and mitigating assurance risks.

The current version of the RDP-AIA (v1.1) is largely SME-driven and manual. Future versions may seek to implement this protocol in a self-directed manner (e.g., with reduced need for an interviewer/facilitator), automate aspects of the data collection or artifact generation process, and/or the way in which new cases are added or updated.

2.3 Goal Statement

The goals of the RDP-AIA can be formally stated as:

> *A standardized process designed to*
>
> *a) increase awareness of a wide range of, and the relative tradeoffs between, AIA needs that can inhibit the development of trustworthy AI,*
>
> *b) highlight the potential effects of addressing those needs in minimizing negative and maximizing positive mission-related and/or societal credible impacts that may result from deploying a proposed AI solution, and*
>
> *c) provide a means to build appropriate and well-calibrated levels of trust in deployment and use of the AI solution.*

2.4 Structure of the Document

The remainder of this document is divided into four major sections. Section 3 is focused on preparation for administration of the RDP-AIA with a particular Stakeholder. Section 4 is the largest section of this document and captures the four phases of the interview-based RDP-AIA. The penultimate section (Section 5) is dedicated to collating elicited information into and generating the Priority Report. Follow-up activities are described in the final section (Section 6).

- Section 3: Set-up and Preparation
- Section 4: The RDP-AIA
  - 4.1: **UNDERSTAND** Phase
  - 4.2: **IDENTIFY** Phase
  - 4.3: **DISCOVER** Phase
  - 4.4: **PRIORITIZE** Phase
- Section 5: The RDP-AIA Priority Report
- Section 6: Follow Up

Section 4 follows a common structure, with each sub-section (Sections 4.1-4.4) describing the requirements of the respective phase. These include:

- Description and purpose of the current phase of the protocol
- Required inputs for the current phase (i.e., artifacts produced before entering a specific phase)

- Expected outputs of the current phase (i.e., artifacts produced in this phase).

Sections 4.1-4.4—the four phases of the RDP-AIA—each contain multiple steps that follow a common structure:

- Purpose of the current *step* within each phase
- Questions (prefaced throughout the document by "ASK:") and/or activities (prefaced by "ACTIVITY:") to generate content for the outputs for each step.

QUESTION / ACTIVITY

Throughout this document, words meant to be explicitly spoken by the facilitator during the interview (e.g., participant instructions, questions, activity descriptions) are presented in a light green box (like this one) and denoted with quotation marks. *[Additional instructions are provided in italics and square brackets like here].*

Hint

Interviewer advice, that might be helpful to heed during the interview, is provided throughout this document and presented in a white box (like this one).

- A description of how the information generated during each step should be used in a subsequent step or phase of the RDP-AIA. When not self-evident, the expected output will also be described.

This document also includes several appendices containing the artifacts described in Section 2.1 (Overview). Use of these artifacts will be described, explained and, where appropriate, called out with hyperlinks at appropriate points in the body of this document.

3 Set-Up and Preparation for the Interview

3.1 Instigating Event

Preparation for the RDP-AIA begins when a Stakeholder contacts MITRE's AIAD Lab with specific or general AI assurance concerns. These concerns could be regarding a problem space or particular challenge that may be addressed using AI, or related to a specific proposed solution involving AI.

3.2 Scheduling Email

The first preparatory step is for the AIAD Lab team to send an email to schedule the interview. The initial email should include an overview of the RDP-AIA and ask the Stakeholder to reserve enough time to answer the pre-interview questions, which will be sent separately. An example email can be found in Appendix J (section J.1).

3.3 Pre-Interview Questions

Once the interview has been scheduled, the AIAD Lab team should send a follow-up email containing the pre-interview questions. These questions are asked ahead of the interview to allow the AIAD Lab team to tailor subsequent interview questions as needed and to anticipate which specific AIA needs and past use cases might be relevant to the Stakeholder's effort. The team can modify the list of questions depending on how much information they want to collect ahead of the interview[2]. An example of the pre-interview questions can be found in Appendix J (section J.2), which anticipate several of the questions from the UNDERSTAND phase of the interview.

3.4 Document Collection

As requested in the follow-up email sent to Stakeholders (see Section 3.3), background materials, responses to the questions posed in the email, and any associated documentation should be gathered to help the AIAD Lab team prepare to carry out the RDP-AIA with the Stakeholder. This includes:

- Familiarizing the AIAD Lab team with project-specific terminology and reference concepts
- Improving understanding of the Stakeholder's mission problem and/or AI solution
- Facilitating creation of the Stakeholder's Use Case Description
- Capturing documents that can be attached to this Use Case for added context and later reference by others
- Identifying similar historical cases with similar mission problems or AI solutions

[2] Note that answers to some of the pre-interview questions act as input to the Use Case Description, which should be partially completed prior to Phase 1: UNDERSTAND. Some of the pre-interview and UNDERSTAND questions are purposely similar to permit the AIAD Lab team to seek any necessary clarification during Phase 1.

Overall, the effect of document collection should be to make the interview discussion more focused and engaging and to broaden the AIAD Lab team's understanding of the system to enable discussion of associated AIA needs.

3.5 Other Preparation

Based on activities so far, the team may prepare for the interview in the following ways:

- Tentatively filling in parts of the Use Case Description (see UNDERSTAND Phase).
- Taking a first pass at identifying some analogous past use cases (see DISCOVER Step 3).
- Identifying responses from the pre-interview questions that need clarification.

4 The Risk Discovery Protocol for AI Assurance

To enable the AIAD Lab Team to synthesize information elicited during one phase in preparation for another, we recommend that the 4-phase RDP-AIA be implemented in a minimum of two, ideally three parts (e.g., Part 1: UNDERSTAND, IDENTIFY; Part 2: DISCOVER; Part 3: PRIORITIZE). Ideally, these would be implemented with sufficient time in between phases/parts to review and synthesize the material generated.

Supplemental materials (e.g., multiple choice response option for specific questions, response scales, definitions, etc.) to support conducting each phase of the protocol are available in the *RDP-AIA-Supplemental Materials* PowerPoint document.

4.1 UNDERSTAND Phase

Description and Purpose:
Phase 1 of the RDP-AIA constitutes the first face-to-face interaction with the Stakeholder. Using a structured discussion/interview-based format, the primary goal is to build a comprehensive understanding of the Stakeholder's mission problem, proposed AI solution, and expected effects. The aim is to obtain sufficient contextual information during the UNDERSTAND phase to engage in a meaningful discussion with the Stakeholder in subsequent phases of the protocol that permits them to identify, discover, and prioritize their respective AIA needs.

Required Input:

- Partially Completed Use Case Description

This phase requires input from the Stakeholder's responses to the pre-interview questions and documents collected during interview preparation (see Sections 3.3-3.4). Before commencing this phase, this information should be collated and used to partially populate the Use Case Description (see Appendix A).

Hint
Throughout the UNDERSTAND phase, when relatively complete information is provided in response to pre-interview questions and data collection, the interview team should use questions and activities in this phase as an opportunity to confirm, clarify, and/or elaborate on the information provided.

Expected Output:

- Completed Use Case Description

The primary output of this phase is a completed Use Case Description that captures the AIAD Lab team's understanding of the mission problem and proposed solution and characterizes each according to certain features. The Use Case Description can be found as part of the Priority Report (Appendix I), with detailed instructions in Appendix A.

4.1.1 UNDERSTAND Step 1: Identify the Mission Problem

The goal of this step is to understand the mission problem that the Stakeholder's solution addresses, to unpack the specific challenges that problem poses, understand why (and/or how) these are challenging and, ultimately, determine what, exactly, their solution is designed to achieve. One of the key aims of this step is to introduce key concepts and provide the AIAD Lab team (and stakeholders) with a common terminology and frame of reference specific to the problem.

QUESTION

ASK: **"**Please can you briefly summarize the mission problem or specific challenge your system addresses (i.e., what is it trying to achieve)?"

Hint

Encourage the Stakeholder to focus on the problem / challenge being solved, overall strategic intent, and/or functional purpose (i.e., the ends) that the AI solution will address rather than the specific tactics/tasks/actions the solution will perform (i.e., ways/means) to solve the problem/achieve intent. Questions about the Solution itself are asked below.

ACTIVITY

*[**INSTRUCTION:** Once the mission problem has been elicited, **RESTATE** this back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify]*

How should the information be used? Answers to this question should inform completion of the description of the *Mission Problem* in the *Use Case Description* (See Appendix A).

4.1.2 UNDERSTAND Step 2: Identify the Sector

The goal of this step is to identify the domain/sector (such as healthcare, national security, intelligence, etc.)—from the list below[3]—so that it can be properly characterized and, where relevant, compared to other cases in the same domain or sector.

QUESTION

*[**INSTRUCTION:** Using the Support Materials, present the list of sectors to Stakeholders. This list is repeated below for reference)]*

ASK: **"**What is the use case domain, sector, or industry?"

SAY: "Please select a Sector from the list."

[3] This list was adapted from the UN Classification of the functions of government. See Appendix F for more details on how this list was derived.

Sector List[4]:

1. National security
2. Intelligence and information operations
3. Law enforcement, judicial, prisons
4. Emergency management
5. Health and wellness
6. Administration and finance
7. Transportation
8. Industrial and manufacturing
9. Environment, energy, agriculture
10. Communications
11. Space
12. Education, recreation, culture
13. Housing and community amenities
14. Social protections (equity, unemployment)
15. Other (Please Specify)
16. Domain-agnostic

How should the information be used? Use the identified sector to inform the entry in the *Sector* row of the Use Case Description.

4.1.3 UNDERSTAND Step 3: Identify the AI Application Area

The goal of this step is to understand the function or capability of the AI solution (such as visual classifier) and identify the appropriate classification from the list below.[5] Identification should permit the solution to be properly characterized and, when added to the Use Case Repository, compared to other use cases with similar characteristics.

QUESTION

*[**INSTRUCTION:** Using the Support Materials, present the AI Application Area list to Stakeholders. This list is repeated below for reference)]*

ASK: "What is the general application area of the AI system, or what AI functions is it expected to perform?"

SAY: **"**Please select an AI Application Area from the list AND then generate a specific description of the application area."

AI Application Area list:

1. Computer Vision
2. Natural Language Processing
3. Planning & Scheduling
4. Robotics & Autonomous Systems
5. Decision Aids & Predictive Analysis
6. Other (Please Specify)

[4] All materials required for administration of the RDP-AIA are available in PPT form (see Materials for the RDP-AIA).

[5] This method of classification of AI solutions was proposed in a recent MITRE report on future T&E for AI (Reeder et al., 2022). See Appendix A for a text description.

Hint

Specific descriptions should not be limited to a given taxonomy, but follow generally understood application domain names (e.g., Computer Vision / Image Classification; Natural Language Processing / Topic Modeling)

How should the information be used? Use the identified AI function/capability to inform the *AI Application Area* row of the Use Case Description.

4.1.4 UNDERSTAND Step 4: Identify the Maturity Level

The goal of this step is to clarify the current development status of the AI solution in question. Note that stakeholders and sponsors are likely to have different assurance concerns and priorities in early development, late development, and deployment[6]. Changing priorities across the lifecycle is addressed throughout the RDP-AIA.

QUESTION

ASK: "Now let's talk about your solution. Which of the following categories best describes your technology readiness or maturity level: Early Development, Late Development, or Deployed?"

Hint

Solutions not yet in development should be considered *Early Development*. If desired, the stakeholder can specifically place their system on the 9-point Technology Readiness Scale (TRL) scale used by the DOD.

How should the information be used? Use the identified stage of development/deployment to inform the *Maturity* row of the Use Case Description.

4.1.5 UNDERSTAND Step 5: Unpack the Solution Workflow

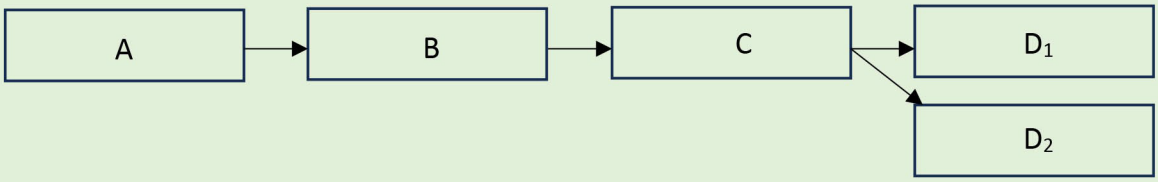
The goal of this step is to gain an understanding of the sequence of key steps/tasks and/or technical components in the Stakeholder's solution, and to capture that information in a simplified AI solution workflow. This activity should also identify references to data collection and storage, algorithms, inputs, and outputs, and expected impacts. Human touchpoints with the workflow will be elaborated in a subsequent step.

A key aim of this step is to identify key terminology, concepts, and operations in use in the Stakeholder's solution and to provide the AIAD Lab team with a common frame of reference specific to the solution.

[6] This question and related questions in the subsequent phase (IDENTIFY) are focused on understanding Stakeholders' *current* assurance needs based on their current stage of development/deployment. In the final phase (PRIORITIZE), Stakeholders will be asked to prioritize their assurance needs for both the current stage of deployment AND post-deployment.

ACTIVITY

SAY: "In more than 3- but less than 6-steps, please summarize the key sequence of activities or components in your AI solution workflow, e.g.,"



Hint

Encourage the Stakeholder to describe the solution workflow in tactical rather than strategic terms, and describe how each step/component supports mission task performance.

ACTIVITY

*[**INSTRUCTION:** Once the task sequence has been elicited, **RESTATE** this back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify]*

How should the information be used? Use a summary of the solution workflow (see above Activity Box/task diagram) as an entry into the *Solution Workflow* row of the Use Case Description. The summary should list/describe the steps and/or components of the system and provide a rough outline of how they are connected.

4.1.6 UNDERSTAND Step 6: Identify the Algorithm Type(s)

The goal of this step is to identify the type(s) of algorithms—from the list below[7]—used in the AI solution to solve the mission problem. This will allow technically similar use cases (i.e., those employing the same or related algorithms) to be identified and compared with the current use case (see Appendix C for how to identify similar use cases).

QUESTION

ASK: "Which method(s) or algorithm(s) is your AI solution using?"

SAY: **"**Please select from the hierarchy and then provide a specific description of the AI method/algorithm"

AI Method/Algorithm List

- Knowledge Representation & Reasoning
- Machine Learning – Supervised

[7] This method of classification of AI methods/algorithms was proposed in a recent MITRE report on future T&E for AI (Reeder et al., 2022). See Appendix A for a text description.

- Machine Learning – Unsupervised
- Machine Learning – Semi-supervised
- Machine Learning – Reinforcement Learning
- Other (Please Specify)

Hint

Any solution utilizing symbolic AI and dependent (to some extent) on explicit domain knowledge (i.e., not machine learning), can be classified as Knowledge Representation & Reasoning, if the Stakeholder agrees. This includes, for example, state machines, evolutionary algorithms, and goal-based planners, as well as expert systems.

Hint

Specific descriptions should not be limited to a given taxonomy, but follow generally accepted method/algorithm names (e.g., Knowledge Representation & Reasoning / Stochastic Rule System; Machine Learning – Supervised / Convolutional Neural Network)

How should the information be used? Use the selected method/algorithm and specific description of the same to inform completion of the *Algorithm / Method* row of the Use Case Description.

4.1.7 UNDERSTAND Step 7: Identify Human Touchpoints

The goal of this step is to identify important human-AI interdependencies in the solution workflow, including points in the workflow where humans interact, or should be able to interact, with the system.

QUESTION

ASK: "Where and how are humans involved in the AI solution workflow? Where are humans 'in,' 'on,' or 'out' of the loop? What are the human-AI interdependencies?"

Hint

Interactions may be in the form of a single point of contact (if the human simply receives the system output) or multiple points of contact (e.g., if the system accepts feedback while performing the task or requires user other input). Likewise, the humans and/or AI tasks in the workflow may be independent of or dependent on the other entity, or interdependent.

ACTIVITY

*[**INSTRUCTION:** Once the human touchpoints have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify]*

How should the information be used? Use a bulleted list of human-system interactions and interdependencies generated to inform completion of the *Human Interaction* row of the Use Case Description.

4.1.8 UNDERSTAND Step 8: Identify the Expected Effects

The goal of this step is to identify the anticipated effects and impacts of deploying the system if all goes as intended.

QUESTION

ASK: "If your system is deployed and everything goes right, what would you expect to happen? What are the expected outcomes, mission impacts, or societal benefits?"

Hint

Expected effects can include both low-level positive effects or outcomes (e.g., "the classifier can classify incoming images at 95% accuracy") and high-level effects (e.g., "fatalities are reduced"). It may also include negative effects where those are already anticipated (e.g., "additional training will be needed for initial adoption to succeed").

ACTIVITY

*[**INSTRUCTION:** Once the outcomes/impacts have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify]*

How should the information be used? Use the summary of expected outcome of using the system, or a list of the same, to inform completion of the *Expected Effects* row of the Use Case Description.

4.2 IDENTIFY Phase

Description and Purpose:
The primary goal of the second phase of the RDP-AIA is to (a) elicit the specific AIA needs and/or concerns Stakeholders (or their Sponsors) would like to address and (b) understand the basis for the specific concerns raised, including the potential impacts (across the lifecycle) of not addressing these needs.

Required Input:

- Use Case Description

The Use Case Description generated in the UNDERSTAND phase allows the interview team to tailor and elaborate questions in the IDENTIFY phase to the Stakeholder's mission problem and AI solution.

Expected Output:

- AIA Needs List with Justifications (See G.1 Working with AIA Needs)

The primary output of this phase is an unstructured list of AIA needs elicited from the Stakeholder and an associated rationale or justification for those needs. This output should provide sufficient information to permit the interview team to understand the primary reasons for a particular need or concern. The final AIA Needs list can be found as part of the Priority Report (Appendix I). If helpful, the AIAD Lab team can utilize the aids in Appendix G to work with AIA needs before pasting them into the Priority Report.

4.2.1 IDENTIFY Step 1: Elicit Stakeholder AI Assurance Needs

The goal of this step is to identify needs or concerns that have been explicitly stated or assumed to be priorities by the Stakeholder and/or sponsor. These could come from firsthand experience (e.g., personal observation) or from other sources (e.g., research by others).

QUESTION
ASK: "What are the specific-Sponsor assurance needs that you are most concerned about, and why?"

Hint
When answering the "why" question, choose the justification that best 'fits' the rationale provided from the source list below AND then add a more specific description of the reason provided, for example, *"Lesson Learned: Previous implementation failed to be adopted, in part, due to this".*

Justification (Why?) Source List:

1. **Observed**: This need has been observed *directly* by you during operations, research, or experimentation

2. **Lesson Learned**: This need was determined from an in/formal incident review (e.g., AAR, hotwash).
3. **Inferred**: This need is assumed or intuited from experience without drawing on specific observations.
4. **Policy**: This need is specified in policy.
5. **Authority:** This need was determined by a higher authority (e.g., commander) but the Stakeholder doesn't necessarily know the reason.
6. **Tradition:** This need was determined based on common or typical practice (e.g., this is what is always done).
7. **Research**: General research findings (by others) indicate or imply this is a need.

ACTIVITY

*[**INSTRUCTION:** Once the needs have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify the needs and associated rationale.]*

ASK: "I heard you say that *X* is a priority because *Y*. Is that right?"

How should the information be used? Add these needs and justifications to the AIA Needs list.

4.2.2 IDENTIFY Step 2: Identify Any Resolved Assurance Concerns

The goal of this step is to identify needs that have already been addressed and, therefore, are no longer a concern. Such needs may not have been mentioned in response to the previous question given their prior resolution. These concerns should be captured along with their resolutions because (1) they may be helpful to other Stakeholders with similar concerns and (2) they may become relevant again as the project advances.

QUESTION

ASK: "Are there any other assurance needs that you've already addressed? If so, what were these and how did you address them?"

ACTIVITY

*[**INSTRUCTION:** Once the needs have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify.]*

ASK: "I heard you say that you addressed *X* by doing *Y*. Is that right?"

Hint

The stakeholder should be able to explain why the need is no longer a priority. If actions were taken, record them and their outcomes for other teams that might have similar concerns. If no action was taken, why did priorities change?

How should the information be used? Add these needs and justifications to the AIA Needs list. In the Notes column, be sure to capture why the need is not currently a concern.

4.2.3 IDENTIFY Step 3: Identify Potential Future / Scaling Issues

While the previous questions covered current and past concerns, the goal of this step is to anticipate future concerns. Concerns that are not currently on the Stakeholder's mind may become increasingly important as the project advances toward deployment. Potentially, past concerns could continue or resurface. Future concerns should be accounted for ahead of time.

QUESTION

ASK: "Post deployment, if the system functions as intended, what assurance issues might remain or arise as the use of the system broadens, or as the real-world starts to diverge from your initial expectations? Which factors might contribute to or lead to these issues?"

ACTIVITY

*[**INSTRUCTION:** Once the needs have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify.]*

ASK: "I heard you say that *X* could be a concern because *Y*. Is that right?"

Hint

Try to question any assumptions the Stakeholder might be making during development—especially those that may not necessarily hold post deployment. You may also use this step to explore differences in expectations between the testing and real-world environment, or post-deployment changes in the user, adversary, context, or changes over time.

How should the information be used? Add these needs and justifications to the AIA Needs list. In the Notes column, be sure to capture that the need is a future concern.

4.2.4 IDENTIFY Step 4: Identify Any Other Concerns

The goal of this step is to provide an opportunity to capture any remaining AIA assurance concerns not yet discussed.

QUESTION

ASK: "Are there any other assurance needs you haven't yet mentioned that are currently on your radar with respect to this project? If so, what are they and why are they of interest?"

ACTIVITY

*[**INSTRUCTION:** Once the needs have been elicited, **RESTATE** these back to the Stakeholder—using their own words where possible—to permit them to verify and/or clarify.]*

ASK: "I heard you say that *X* could be a concern because *Y*. Is that right?"

How should the information be used? Add these needs and justifications to the AIA Needs list.

4.3 DISCOVER Phase

Description and Purpose:
The primary goals of phase 3 of the RDP-AIA are to (a) translate the unstructured AIA needs and concerns (from IDENTIFY) into a defined set of AIA needs categories and specific needs, and (b) explore the broader landscape of potential AIA needs to help further qualify Stakeholder needs and/or permit discovery of unanticipated risks. The aims of DISCOVER are to facilitate Stakeholder understanding of the impact of AIA needs on mission-related and sociotechnical outcomes and the ability to develop trustworthy AI, and to highlight how (not) assuring these outcomes can impact human trust.

Required Input:

- AIA Needs List with Justifications (See G.1 Working with AIA Needs)
- AIA landscape (Appendix D)
- Glossary (Appendix E)
- List of Similar Cases (See Appendix C)

In cases where sufficient numbers of Use Cases already exist, *prior to commencing the DISCOVER phase* the interview team will query the Use Case Repository and generate a list of past use cases similar to the current use case[8] (see Appendix C for a description of the process for querying the Use Case Repository and identifying similar use cases). This phase also requires access to the unstructured list of Stakeholder's AIA needs and concerns generated during IDENTIFY, as well as access to the AIA landscape (see Appendix A) and Glossary (see Appendix E), which contains a broad set of AIA specific needs organized by needs category.

Expected Output:

- Updated AIA Needs List with Justifications

The primary output of this phase is an updated list of AIA specific needs (and associated AIA need categories) extracted from the AIA landscape, along with justifications. Some of the specific needs and justifications will come from the initial AIA Needs List with Justifications (see IDENTIFY: Expected output). Additional AIA needs and their associated justifications will be elicited during DISCOVER. The final AIA Needs List can be found as part of the Priority Report (Appendix I). If helpful, the AIAD Lab team can utilize the aids in Appendix G to work with AIA needs before pasting them into the Priority Report.

4.3.1 DISCOVER Step 1: Introduce the AIA Landscape

The goal of this step is to ensure the Stakeholder understands the AIA Landscape (including its origin) and its organization (e.g., need categories > specific needs) sufficiently well to be able to identify their assurance needs within it.

[8] Use case entries in the Use Case Repository will consist of multiple (a) Use Case Descriptions from past cases, and (b) RDP-AIA Priority Reports corresponding to the respective Use Case Descriptions. These artifacts permit an assessment of similarity between the current interview and past cases (that have completed the RDP-AIA) and allow AIA priorities from similar past cases to be leveraged in the current interview (see Appendix C for more details).

ACTIVITY

*[**INSTRUCTION**: Present the AIA Landscape (Appendix B) and describe it using the following wording:]*

SAY: **"**This is a representation of the AI assurance landscape that we generated. It was created using both Stakeholder input and from the many different AI Assurance frameworks and guidelines that have been published to date."

SAY: **"**A few things to note:

- Each Stakeholder and each published document uses a different framework.
- Each one emphasizes different aspects of assurance.
- Many use different terminology to talk about similar assurance concerns.
- Others use the same term to refer to different needs.
- While there are some obvious differences across frameworks, there are also many similarities.

Our goal in compiling this landscape was to create a domain-agnostic taxonomy that captures the nuances and similarities across frameworks and domains. This is a work in progress and will likely be updated over time as we learn more."

***ACTIVITY** (Continuation)*

*[**INSTRUCTION**: Walk through the AIA Landscape using the following wording:]*

SAY: **"**Let's walk through the landscape. At the top we have *Trustworthy AI*, defined as 'operating within acceptable level of risk'. Underneath that there are 11 needs categories, each containing a set of specific needs.

We'll go through each one as we ask further questions. Note that some specific needs (like transparency) appear in multiple categories—to denote the different ways in which these concepts are used by different Stakeholders. Where this is the case, a qualifier has been added to the respective definition in the Glossary.

The underlying premise is that assuring or not assuring each need can have positive and negative impacts on the mission and society more broadly."

***ACTIVITY** (Continuation)*

*[**INSTRUCTION**: Define each of AIA Landscape 'categories' one by one using the Glossary (reproduced below for convenience)]*

SAY: "Next, I'll briefly define each category and highlight the specific needs under each. Let's start with Integrity-Enabled, which is defined as…"

Assurance Needs Category	Definition
Integrity-Enabled	Satisfies expectations for technical and scientific integrity
Effective	Achieves intent and/or desired outcomes
Secure	Resistant to unauthorized activities
Governable	Implements a framework of policies, rules, and processes for appropriate oversight within and across relevant organizations
Safe	Does not lead to a state in which human life, health, property, or the environment is endangered
Accountable	Answerable to the stakeholders it empowers and to those it impacts for its proper and appropriate functioning, and obligated to address identified deficiencies
Private	Safeguards information collection and use to preserve autonomy and dignity
Interpretable	Makes processes and outputs apparent and meaningful in the context of functional and anticipated purposes
Equitable	Addresses disparities in use and outcomes across individuals and groups
Human-Empowered	Leverages human capabilities and enables pursuit of human goals
Civil	Designed and operates in accordance with social norms and the public good

Hint

Where necessary, use the definitions of AIA 'specific needs' (cf. needs categories) (see Glossary; Appendix E) to aid Stakeholder clarification. The next step will be to locate the Stakeholder's specific AIA Needs (generated in IDENTIFY) within the AIA Landscape where they will identify the AIA Landscape need categories and specific needs that correspond with their stated needs.

How should the information be used? No specific output is expected to be generated from this step.[9]

4.3.2 DISCOVER Step 2: Locate Needs in AIA Landscape

The goal of this step is to locate Stakeholder's AIA needs within the AIA Landscape.

[9] In the event that the Stakeholder identifies AIA needs (categories or specific needs) not captured in the AIA Landscape, this information should be captured and used in periodic review/updating of the AIA Landscape/Glossary.

QUESTION

ASK: **"**First, we want to locate your assurance needs in this AIA landscape. Which of the AIA specific needs do you think most closely captures the specific Sponsor assurance needs you've mentioned so far?"

Hint

The team may choose to have the stakeholder verbalize the relevant needs (and associated categories), tag words on a collaborative platform such as Mural, print out and circle words manually, or even have the stakeholder browse and pick out words from the Glossary.

ACTIVITY

*[**INSTRUCTION:** Check the AIA Needs List (generated from IDENTIFY) to confirm all AIA needs categories and specific needs have been located in the AIA Landscape].*

Hint

Repeat the process above with the aid of the Glossary, to walk the stakeholder through any specific needs not yet located on (or translated in terms of) the landscape. Ask for clarification as needed.

Note. Make sure at least one 'specific need' on the AIA Landscape is identified for each of their stated assurance needs (from IDENTIFY), especially if the Stakeholder talks about their assurance needs in 'category' rather than 'specific need' terms.

How should the information be used? Specific needs (and corresponding needs categories) identified in the AIA landscape should be used as input to the Updated AIA Needs List with Justifications.

4.3.3 DISCOVER Step 3: Explore Similar Cases[10]

The goal of this step is to use AIA priorities identified in similar past use cases as a stimulus for exploration and discover of potential assurance needs not yet discussed. Prior to commencing this step, the AIAD Lab team should generate a list of similar use cases via the process described in Appendix C. The aim is to identify past cases in the Use Case Repository that are technically similar and explore the respective AIA priorities.

[10] Until multiple (e.g., >25) use cases have been generated through engagement with Stakeholders (via the RDP-AIA), and until similar past use cases can be identified using more advanced methods described in Appendix C (see Method 3 & 4), the AIAD Lab team should use their experience to identify similar cases via either manual or category-based retrieval (see Appendix C: Method 1 & 2).

QUESTION

[INSTRUCTION: *One by one present a similar past use case (max. 3) extracted from the Use Case Repository. Highlight their basis for similarity with the current use case (see Appendix C), and the AIA priorities from the past case].*

ASK: **"**Based on previous lessons learned from use cases similar to yours, we've identified other potential assurance needs. How relevant are these cases and the associated assurance needs to your situation?"

Hint

For each similar use case extracted from the Use Case Repository, explain the basis for similarity between the past and current case (e.g., same domain, algorithm, similar mission problem) and, where appropriate, provide relevant context from the past case (from its Use Case Description or corresponding Priority Report).

Note. To avoid overload, present no more than THREE similar use cases.

How should the information be used? Add additional specific AIA needs identified from the discussion of similar cases to the *Updated AIA Needs List with Justification.* Append additional specific needs generated in this step with the respective identifier for the similar Use Case. Add the respective identifier to the *Analogous Case* row of the Use Case Description in the Priority Report.

4.3.4 DISCOVER Step 4: Explore Other AIA Needs

The goal of this step is to explore other, potentially related AIA needs not yet discussed and to identify whether they pose any concern to the Stakeholder and/or are of relevance to their mission.

QUESTION

SAY: **"**Across all stakeholders, some have prioritized assurance needs that "you" didn't mention, such as those captured by some of the other categories in the AI Assurance landscape. We want to understand whether any of these specific needs present an assurance concern for your particular mission problem or proposed solution.

ASK: Let's start with *Integrity-Enabled*. Given your current stage of development, do you have any concerns about any of the 'specific needs' that we haven't discussed under this category?"

List of AIA Landscape Needs Categories:

- Integrity-Enabled
- Effective
- Secure
- Governable
- Safe
- Accountable

- Private
- Interpretable
- Equitable
- Human-Empowered
- Civil

QUESTION (If not already stated)

ASK: "Why are you concerned about [specific need X]?"

Hint

Have the Glossary (Appendix C) to hand, provide definitions of assurance needs categories and specific needs and clarify any ambiguities as required.

- IF a specific need is mentioned as a potential concern (i.e., one that is not yet on the *Updated AIA Needs List with Justification*), THEN add it to the list.
- IF the stakeholder selects a specific need that appears in more than one category, THEN ask the stakeholder to review all instances of that need in other categories.
- IF the specific needs listed in a category are NOT a concern, explore if any are *relevant to* the Mission Problem or Proposed Solution (rather than being particularly concerning), before moving to the next category.

Note. The team may decide to utilize an alternate view of the landscape for this step, such as those in Appendix H. These kinds of views would be most helpful when used on a large scrollable platform such as Mural so that the stakeholder could read and navigate them.

ACTIVITY

*[**INSTRUCTION**: Repeat the question, replacing last AIA need category (e.g., Tech/Enabled) with the next in the list of AIA Landscape categories (see below)].*

QUESTION

ASK: "Are there any other assurance needs not covered by the AIA Landscape?"

How should the information be used? Add additional AIA needs to the *Updated AIA Needs List with Justification*. Include in the justification, any 'connected' needs (e.g., justification based on links to other specific needs).[11]

[11] Future techniques for exploring additional connections between specific needs could include word embeddings and LLMs.

4.4 PRIORITIZE Phase

Description and Purpose:
The primary goal of phase 4 of the RDP-AIA is to prioritize the AIA needs already identified. This is accomplished by assessing current (and future) risks, and prioritizing assurance needs based on the risks relative to assessed risk tolerance levels.

Required Input:

- Complete/Updated AIA Needs List with Justifications

The updated list of complete AIA Needs (together with their justification) is the primary input for PRIORITIZE.

Expected Output:

- Risk Assessment Cards
- List(s) of AIA Needs, prioritized by Risk & Risk Tolerance Levels, and associated justifications (e.g., As-is vs. At-Deployment).

The output of this phase is a set of Risk Assessment Cards (see Appendix G3) and TWO prioritized list(s) of AIA needs. The Risk Assessment Cards capture the impacts and the severity/likelihood, risk, and risk tolerance levels for each specific assurance need.

The first prioritized list is a rank-ordered (highest to lowest) list of priority needs based on the system "as-is" (i.e., at the current stage of development). The second is a rank-ordered (highest to lowest) list of priority needs based on the system "at deployment". The second list of needs will be generated if/when assurance needs, impacts, and associated risks differ "at deployment" from "as-is" (see PRIORITIZE Step 4).

Each priority list includes a range of corresponding information—captured in the Risk Assessment Cards—including the worst credible negative impact, severity and likelihood of those impacts, derived risk levels and assessed risk tolerance levels, and any associated justifications in narrative form. Each of these data serve as input to the Priority Report.

4.4.1 PRIORITIZE Step 1: Rank-Order AIA Needs

NOTE: If, fewer than 10 specific needs were generated in the DISCOVER phase, this step can be omitted.

This purpose of this step is to constrain the list of AIA needs generated in DISCOVER to a manageable number (e.g., ~6 specific needs) prior to conducting the risk assessment (see PRIORITIZE Steps 2-4). This is achieved in collaboration with the Stakeholder by rank-ordering the complete/updated AIA Needs, in order of concern, from most to least concerning, and pruning this list to <10 specific needs. The AIAD Lab team may refer to the interview aids and examples in Appendix G to facilitate this process.

ACTIVITY

SAY: "Given the specific assurance needs you've identified so far, and our discussion of other potentially relevant assurance needs, please rank order your top 6 or so specific needs (not needs categories) from most concerning [1] to least concerning [~6] for your AI solution.

SAY: Please base your rankings on your CURRENT stage of development (i.e., current TRL, not post-deployment [unless already deployed])."

Hint

Remind interviewees to rank specific needs rather than needs categories. Encourage them to consider factors such as:

- *the effects of not assuring [X]*
- *the relative urgency of [X]*
- *the relevance of [X] for the current stage of development (cf. deployment)*
- *the availability of resources required to assure [X]*
- *the extent to which assuring [X] might trade-off with other specific needs*

How should the information be used? Use the rank-ordered list of ~6 AIA needs as input to the next step of the PRIORITIZE phase.

4.4.2 PRIORITIZE Step 2: Assess "As-Is" Risk

The goal of this step is to assess the risks associated with not assuring the solution "as-is." Risk is assessed, in turn, for each of the identified specific AIA needs. (***Note.*** As-is judgements may differ from judgments about the risks post deployment, which will be assessed later).

ACTIVITY

[INSTRUCTIONS: *Provide definitions of risk and severity (see below)]*

SAY: "Next, I want to familiarize you with the definitions and ratings scales we'll use to assess risk. Here are the definitions. Take a moment to familiarize yourself with them."

Term	Definition
Risk	The potential for negative impact (assessed through a combination of severity and likelihood of negative impact).
Severity	The magnitude of negative impacts on assets, operations, individuals, organizations, the Nation, or society more broadly (NIST 800-30, Table H-3),
Likelihood	The expected frequency of negative impacts, assuming the envisioned scale, frequency, and duration of solution use (NIST 800-30, Table G-3)

ACTIVITY

***[INSTRUCTIONS:** Provide the rating scale for severity of impact].*

SAY: "Here is the rating scale for severity of impact. It ranges from very low—or negligible—to very high—or multiple severe or catastrophic—negative impacts on assets, operations, individuals, organizations, the Nation, or society more broadly. Take a moment to familiarize yourself with the scale."

Qualitative Values	Description of Severity
Very High	Multiple severe or catastrophic negative impacts
High	Severe or catastrophic negative impacts (major damage, loss, or harm)
Moderate	Serious negative impacts (significant damage, loss, or harm)
Low	Limited negative impacts (minor damage, loss, or harm)
Very Low	Negligible negative impacts

ACTIVITY

***[INSTRUCTIONS:** Provide the interviewee with the TWO rating scales for likelihood of impact].*

SAY: "Here are TWO scales for assessing likelihood of impact. ***Scale 1*** defines likelihood in terms of expected frequency of negative impacts in # times per year. ***Scale 2*** defines likelihood in terms of expected frequency of negative impacts per single use of the solution.

When we begin assessing risk, I'll ask you to make judgments of likelihood using just ONE of these likelihood scales. Do you prefer a particular scale? Please choose the one you'll use now."

Scale 1: Likelihood (expressed in frequency of negative impacts per year):

Qualitative Values	Description of Likelihood (per year)
Very High	More than 100 times per year
High	Between 10-100 times per year
Moderate	Between 1-10 times per year
Low	Less than once per year but more than once every 10 years
Very Low	Less than once every 10 years

Scale 2: Likelihood (expressed in frequency of negative impacts per solution use):

Qualitative Values	Description of Likelihood (per use)
Very High	Once per use
High	Between once per use and once every 10 uses
Moderate	Between once every 10 uses and once every 100 uses
Low	Between once every 100 uses and once every 1,000 uses
Very Low	Less than once every 1,000 uses

Hint

Encourage the interviewee to choose the likelihood scale that makes most sense to them in terms of assessing the likelihood of impact. Discard the other likelihood scale.

ACTIVITY

SAY: "In this activity, you'll rate each of your ranked assurance needs, one by one, in terms of the severity and likelihood of impact if that specific need were not addressed. Let's start with the specific need you ranked as your #1 concern [STATE NAME OF SPECIFIC NEED].

ASK: "What is the worst credible impact of not assuring [AIA need ranked #1]?"

SAY: "On the severity scale provided, please rate the magnitude of this impact—on assets, operations, individuals, organizations, the Nation, or society more broadly"

ASK: "Why did you provide this severity rating?"

SAY: "On the likelihood scale you selected, please rate the expected frequency of the worst credible impact"

ASK: "Why did you provide this likelihood rating?"

ACTIVITY

***[INSTRUCTION:** Present the Risk Matrix to interviewees].*

SAY: "We translated your severity and likelihood ratings into a risk rating based on a risk matrix from NIST's Risk Management Framework" [NIST 800-30, Table I-2.]

Severity

Likelihood	Very Low	Low	Moderate	High	Very High
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low

QUESTION

ASK: "For this specific assurance need [AIA need ranked #1], what would you consider to be an acceptable level of risk (Very Low, Low, Moderate, High, Very High)?"

ACTIVITY

*[**INSTRUCTION**: Repeat the severity/likelihood AND risk tolerance ratings, and ask the why questions, for each of the remaining ranked assurance concerns].*

Hint

Begin by gathering risk assessment data from the highest ranked need, followed by the next highest ranked need, and so on, until all ranked needs have been assessed.

How should the information be used? The worst credible impacts should be entered as part of the corresponding table in the Priority Report (see Appendix I). The AIAD Lab team can refer to the interview aids in Appendix G for working with needs and risk assessments.

4.4.3 PRIORITIZE Step 3: Prioritize "As-Is" AIA Specific Needs

The goal of this step is to generate a prioritized list of As-Is specific needs, largely from the rankings and associated rationale, and from the risk and risk tolerance ratings data. In cases where the acceptable level of risk is the same for all AIA needs *and* the level of risk posed is different across needs, this is equivalent to just ordering the specific needs by risk rating. In most other cases, prioritization should be a collaborative activity, where Stakeholders can indicate their relative priorities, post risk assessment, using their risk/risk tolerance ratings as a guide.

QUESTION

*[**INSTRUCTION**: Present risk ratings for each of the rated AIA specific needs back to the Stakeholder]*

ASK: **"**Here are all the specific needs we just discussed. Considering the assessed risk, stated level of acceptable risk, and reasons given for each, how would you prioritize them (i.e., from highest to lowest priority) **based on your CURRENT stage of development**?

*[**INSTRUCTION**: Collaborate with the stakeholder to prioritize AIA needs based on concern ranking, risk rating, level of risk acceptability, and associated reasons for each of the rank-ordered assurance needs.]*

Hint

The AIAD Lab team can refer to the interview aids in Appendix G for working with needs and risk assessments. In person, this could be done by putting each specific need on a separate notecard (e.g., using Risk Assessment Cards). Virtually, the team could create a space to present and reorder the items, such as in a Microsoft Word document or a Mural workspace. Two possible approaches are:

- The stakeholder orders the items into a list, thinking aloud so the team understands the rationale.
- The team creates an initial ordering, then allows the stakeholder to reorder the list. A recommended initial ordering would be first based on the difference between Risk and Risk Tolerance level (i.e., whether risks are within acceptable levels or exceed

ACTIVITY

*[**INSTRUCTION**: After prioritizing the assurance needs, clarify their rationale for each and deconflict any apparent contradictions in rankings, ratings, and/or rationale provided.]*

How should the information be used? The interview aids in Appendix G provide fields for additional notes when working with needs and risk assessments. The AIAD Lab team should capture the prioritization as well as rationale and other points of interest, which will be used to populate the *Risk "As-Is"* section of the Priority Report.

4.4.4 PRIORITIZE Step 4: Forecast Risk at Deployment

The goal of this step is to forecast the expected risks that could be associated with not assuring the solution "at deployment." [12] Since assurance concerns may differ between the current

[12] If the solution is already deployed, these questions can be omitted since they would have already been addressed in previous activities.

stage of development ("as-is") and at a future point in time when the solution is deployed, the goal is to capture if/how Stakeholder's priorities might change once the solution is being used. Concerns would shift from what can be done during development to what can be adjusted and monitored after development. Rather than re-assess all risks already assessed, the focus of this step is identifying any changes (to the previous step's ratings) that are expected if/when the solution is deployed and in regular use.

ACTIVITY

ASK: "In this penultimate exercise, I want you to fast forward in time to the point where your system has been deployed and is in regular use. In particular, I'd like you to consider whether the end user might have different assurance concerns than you or your sponsor.

How might your AI assurance priorities CHANGE between now and post deployment when the solution is in regular use, and how would you RE-ORDER the list you just generated to reflect post-deployment priorities?

*[**INSTRUCTION:** For EACH of the ranked assurance needs…]*

ASK:

- "Is the worst that could credibly happen (if X assurance need wasn't met) still the same post deployment, or would it change (compared to current stage of development)?
- Would you rate severity the same, or would the credible impact be more or less severe (as the ratings given for the current stage of development)?
- Would you rate likelihood of occurrence the same, or would it be more or less likely?
- Why did your ratings increase/decrease post deployment? (e.g., due to differences in urgency, relevance, etc.?)
- Is the acceptable level of risk (Very Low, Low, Moderate, High, Very High) the same, higher, or lower post deployment?"

*[**INSTRUCTION:** Collaborate with the Stakeholder to reprioritize assurance needs "at deployment," and clarify their rationale for any changes (from "as-is") and deconflict any apparent contradictions in rankings, ratings, and/or rationale provided.]*

Hint

This step (PRIORITIZE step 4) is a synthesis of PRIORITIZE steps 2 and 3, but for "at deployment" risks. Ensure that you check with Stakeholders the final priority order for 'at deployment' assurance needs and clarify any ambiguities.

How should the information be used? The interview aids in Appendix G provide fields for additional notes when working with needs and risk assessments. The AIAD Lab team should capture the prioritization as well as rationale and other points of interest, which will populate the Priority Report tables for Risks "At Deployment".

4.4.5 PRIORITIZE Step 5: Lessons Learned

The goal of this step is twofold. First, it serves as a final check to help uncover any overarching concerns or additional reasons provided for those concerns. Second, it can be used to assist future participants with similar cases.

ACTIVITY

ASK: "Based on your experience, what assurance-related advice/lessons learned would you give to new those developing and deploying a system like yours? What are the things they should do or avoid doing? What should they look out for in the future?"

Hint

If the AIAD Lab team identifies specific needs not previously highlighted, they could decide to amend the concept list generated from the DISCOVER phase. Note that additional specific needs will not be incorporated into the prioritization process.

How should the information be used? The lessons learned will be captured in a free text field of the Priority Report and will provide additional evidence to support the rationale for prioritization.[13]

[13] Additional assurance needs identified in the lessons learned may added to the priority list in the RDP-AIA Priority Report.

5 Compiling the RDP-AIA Priority Report

The primary output of the RDP-AIA is the Priority Report. The Priority Report contains SIX parts :

- Part 1 contains a project summary that captures key aspects of the Stakeholder's use case (derived from the UNDERSTAND Phase).
- Part 2 contains the AIA needs identified as relevant (derived from the IDENTIFY and DISCOVER phases) or, subsequently, as a priority (derived from the PRIORITIZE phase).
- Parts 3-5 contain the worst credible impacts associated with each AIA need, the perceived levels of severity and likelihood of each of those impacts, the corresponding risk levels posed by each impact, and the relative level of risk deemed acceptable (derived from the PRIORITIZE phase).
- Part 6 contains recommendations to the Stakeholder. It should contain a summary of the priority AIA needs, justification for prioritization, how the needs might be addressed, and the potential impact of addressing them (derived from all four phases).

The highlights of the Priority Report template are in Appendix I. A full template with additional guidance and instructions for completing the report is available on request. The steps for filling in the Priority Report are:

1. Add the *Author* information.
2. Fill in the *Project Summary*:
    - Under *Project Information* add the task name, sponsor, and point of contact. If not direct funded work, write "indirect" in the Sponsor field, followed by primary transition target.
    - Under *Use Case Description,* add the information generated in the UNDERSTAND phase. Fill in the Analogous Cases field with any similar cases presented and confirmed by the Stakeholder in DISCOVER Step 3.
3. Fill in the *AI Assurance Needs* section:
    - Copy information from the Initial and Updated AIA Needs list generated in the IDENTIFY & DISCOVER phases (pruned to ~6 items in PRIORITIZE Step 1). If helpful, the AIAD Lab team can use the interview aids in Appendix G to facilitate sorting and copying the list of needs.
        - Add the definition of each specific assurance need form the Glossary, and add relevant interview data (e.g., paraphrased sections of the interview) to provide context and to permit understanding of the particular need in question.
4. For the "As-Is" system:
    - Add the worst credible impacts, in the form of IF-THEN statements using information generated in PRIORITIZE Step 2.
    - Add severity and likelihood ratings for each worst credible impact using information generated in PRIORITIZE Step 2.

- Convert the severity and likelihood ratings into an assessed level of risk using the risk matrix (see also Appendix B)
- Sort the specific AIA needs using information generated in PRIORITIZE Steps 2-3, i.e., based on both Stakeholder input and assessed risk levels (i.e., highest risk at the top), risk tolerance levels (i.e., risk that exceed stated risk tolerance levels at the top), and Stakeholder rationale.

5. If specific needs were identified as different "At Deployment" from "As-Is," repeat the last step (Step 4 above) using information generated in PRIORITIZE Step 4 ("At Deployment").
6. Convert the prioritized list(s) of assurance needs into Recommendations:
   - Articulate the needs in order of priority and the basis for prioritization (e.g., risks are above acceptable level of risk).
   - Articulate what might be done to assure each prioritized specific needs (based on the definitions in the Glossary).
   - Articulate how assuring the prioritized needs would address the worst credible impacts.
7. Fill in the *Lessons Learned* section with information collected in PRIORITIZE Step 5.

In addition to providing recommendations to the Stakeholder, the Priority Report provides input to the Use Case Repository (see Appendix C) to enrich future elicitations about other projects that might share certain features and serves as input to the next step of the AIAD Lab process: Measuring and mitigating assurance needs.

6 Following Up

6.1 Follow-up Email

After completing the interview and filling in the Priority Report, the AIAD Lab team should finalize the RDP-AIA with a follow-up email to the Stakeholder. The email should accomplish the following:

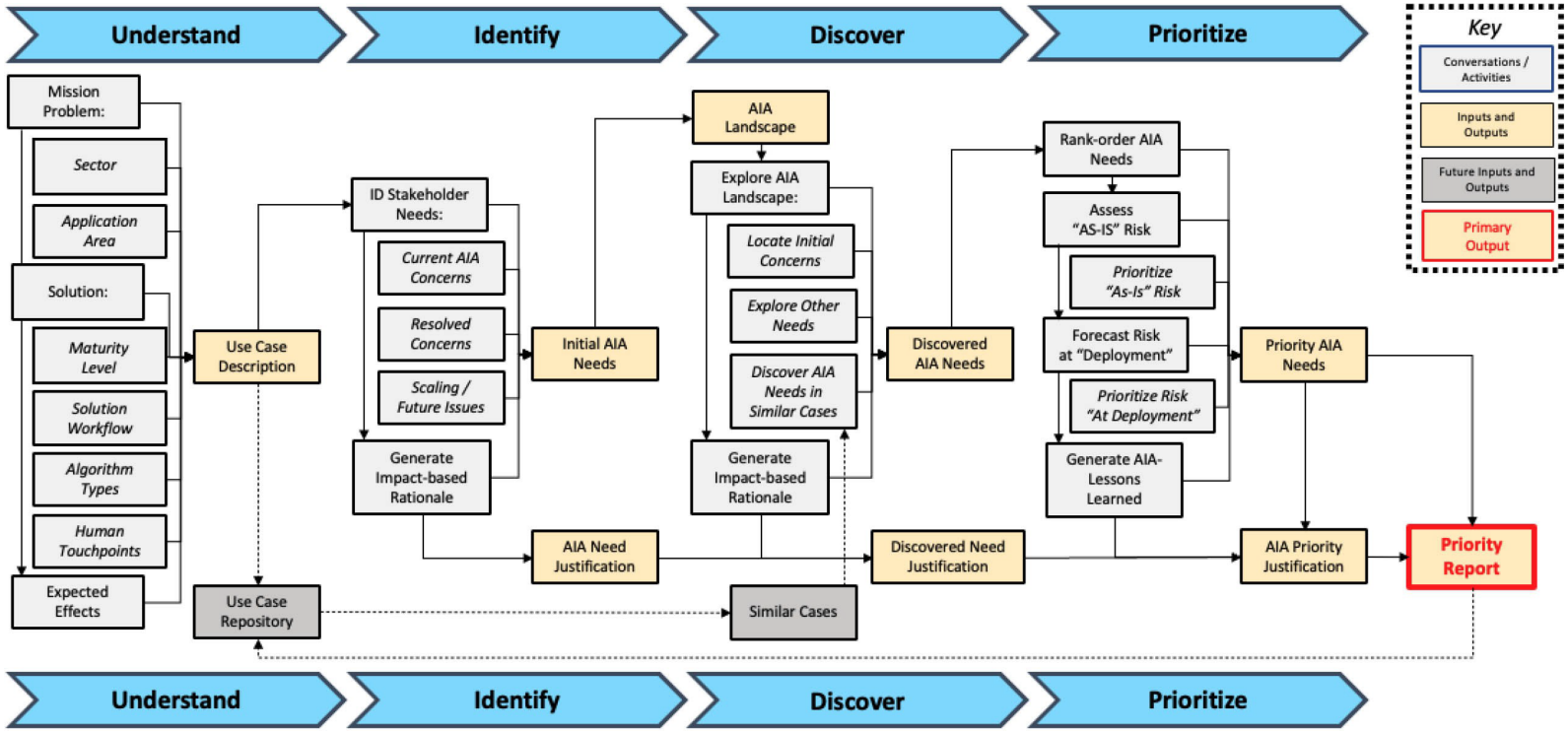
- Confirm the information in the Use Case Description.
- Confirm the prioritized needs, risk ratings, associated rationale, and Recommendations provided in the Priority Report.
- Confirm that the Use Case Description and Priority Report can be shared with other projects and added to the Use Case Repository.
- Serve as a final opportunity for the Stakeholder to bring up any additional needs or lessons learned.

An example email can be found in Appendix J.3.

6.2 Submit the RDP-AIA Priority Report

After finalizing the Priority Report, upload it to the Use Case Repository so it can be referenced in future interviews (see Appendix C), and notify the AIAD Lab team to begin the process of risk measurement and mitigation.

Appendix A: Risk Discovery Protocol for AI Assurance Phases



Appendix B: Use Case Description

This appendix describes the AIA Use Case Description and dimensions/repository fields, and the process for extracting similar/analogous cases.

The Use Case Description can be found as part of the RDP-AIA Priority Report (Appendix I). It is collected during the UNDERSTAND phase, with the goal of creating a useful summarization of how the system is developed and used.

The use case description serves several purposes, including:

- **Summarizing the system:** The use case description should provide a useful summary for reference during the subsequent discussion of risks and mitigations.
- **Retrieving similar past use cases:** The use case description will be used during the DISCOVER phase to find similar previous use cases. Retrieved use cases can be compared to the current use case to suggest other potential risks, or to highlight distinctions that make the current use case's approach unique relative to related approaches.
- **Indexing the use case for later retrieval:** Finally, once the RDP-AIA is complete, the use case and associated materials will be stored in a knowledge base. Just as the use case description was used to retrieve similar use cases, it will itself be made available for retrieval in future instances of the RDP-AIA executed with future Stakeholders.

The Use Case Description from the Priority Report (Appendix I)

Metadata	First created:	*Initial report date*	Last updated:	*Latest date*	Sources:	*Name of point of contact, citation, or other identifier, and how information was obtained*
Problem Description	Mission Problem:	*Task performed or assisted by the AI capability*				
	Sector:	*Multiple choice with write-in*		Application Area:	*Multiple choice with write-in*	
Solution Description	Maturity:	*Early Development, Late, Deployed*		Algorithm Type:	*Multiple choice with write-in*	
	Solution Workflow:	*List the workflow in 3-6 steps*				
	Human Interaction:	*List points of interaction*		Expected Effects:	*List intended effects of using the system*	

The fields in the use case description are as follows, as summarized the following table.

Use Case Field	Field Description	Field Format	Example	UNDERSTAND Step
Metadata				
Created	Date created	Date	2023-08-29	
Last Updated	Date last updated (if different from Created)	Date	2023-08-30	
Source	Source for the use case. May be a point of contact, cited documentation, or other identifiers	Text field	Dr. Jen Doe, jdoe@mitre.org	
Problem				
Mission Problem	The Problem/Challenge the AI capability will help solve	Text field - paragraph	*See main text*	1
Sector	Sector or domain of the application	Domain category	Transportation	2
Application Area	Application area for the AI capability	Superordinate category / Specific category	Computer Vision / Image Classification	3
Solution				
Maturity	Maturity or technology readiness level (TRL) of the solution	Early Development, Late Development, Deployed; or DoD TRL number.	Early Development	4
Solution Workflow	Overview of sequence of steps/tasks performed by the AI capability and/or system components employed to solve the mission problem	Order list of 3-6 sentence-level descriptions	*See main text*	5
Algorithm Type	AI method(s) or algorithm(s) performing the AI capability	Superordinate category / Specific category	Machine Learning – Supervised / Convolutional Neural Network	6
Human Interaction	Interaction patterns and interdependencies between the AI capability and the human user	Sentence-level descriptions of 3-6 interaction points	*See main text*	7
Expected Effects	Intended outcomes of using the system		*See main text*	8

Mission Problem:

This field should contain a single paragraph describing the overall problem that the AI system will be developed to address. It should identify and unpack the specific challenges posed by that problem, understand why (and/or how) these are challenging and, ultimately, to describe the problem their solution is designed to solve.

Sector:

This field should identify the sector in which the application will be deployed. The list of sectors is derived from the UN COFOG government function classification (see Appendix F).

Application Area:

This field should identify the general application area by selecting from the list provided (see left panel of Figure 1) *and* provide a short and specific narrative description of the application domain. This more specific description should follow terminology used to describe generally understood application domains should rather than limited to a given taxonomy (e.g., Computer Vision > Image Classification; Natural Language Processing > Topic Modeling). The top-level application areas are derived from the categories of "AI Application Areas" specified in *Reeder at al. (2022)*[14].

Maturity:

This field should identify the current maturity level of the solution. This can be stated in terms of stage of development/deployment (i.e., early development, late development, deployed) or Technology Readiness Level (TRL, e.g., using the DoD TRL scale). Solutions not yet in development can be labeled Early Development.

Solution Workflow:

This field should contain a single paragraph describing the sequence of key steps/tasks and/or technical components in the Stakeholder's solution and describe how each subsystem supports the completion of the mission task. Where applicable, the description should include references to data collection and storage, algorithms, inputs, and outputs, and expected impacts. It is expected that the 'key' steps and/or 'central' components in most systems can be described in 3-6 steps. For example, a workflow might include an AI classification task, followed by human-in-the-loop validation, followed by retraining based on user feedback.

AI Algorithm Type:

[14] From Reeder, F., Kotras, D., Lockett, J., Pomales, C., & Lokas, R. (2022). *The Future State of Test and Evaluation of Artificial Intelligence - Enabled Systems in the Department of Defense.* MITRE Technical Report, MITRE Corporation, Mclean, Virginia.

This field should identify the general AI algorithm or AI method type by selecting from the list provided (see right panel of Figure 1) *and* provide a short and specific narrative description of the algorithm type (e.g., Knowledge Representation & Reasoning > Stochastic Rule System; Machine Learning – Supervised / Convolutional Neural Network). The top-level algorithm types are derived from the categories of "AI Approaches and Algorithms" specified in *Reeder at al. (2022)*.

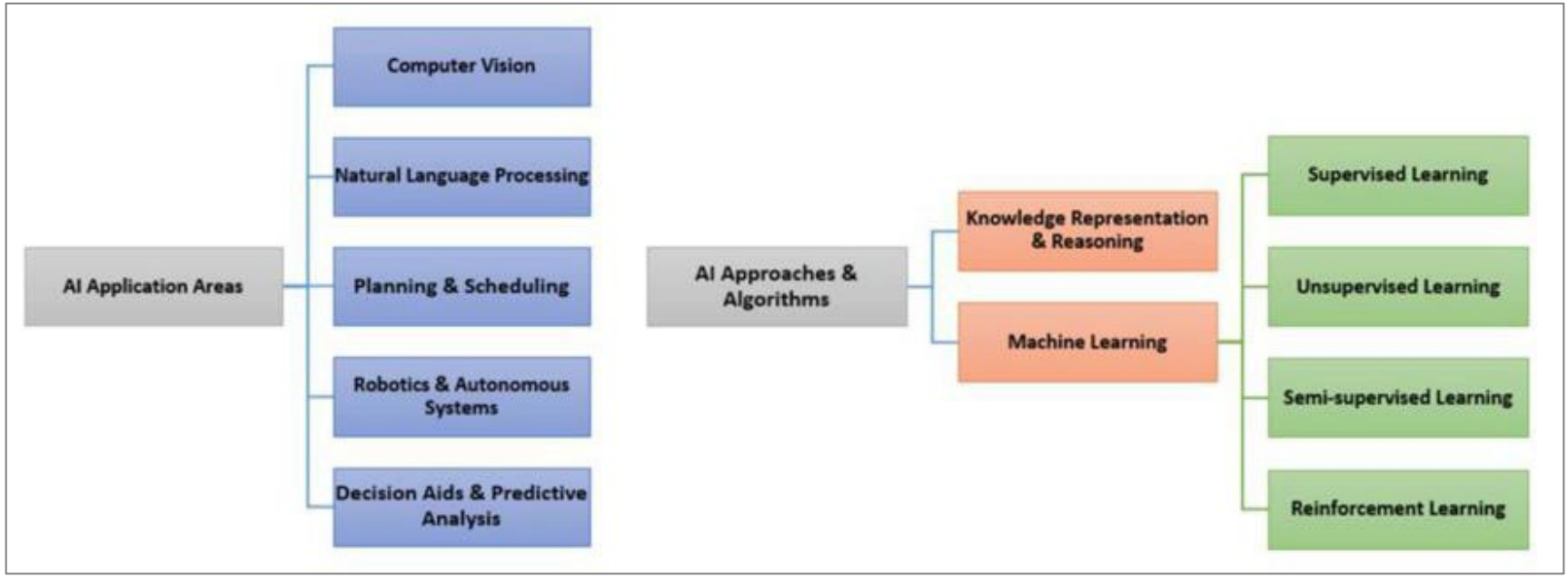


Figure 1: AI Application Areas and Algorithm Types used for the Use Case Description. These general types are used to categorize the overall AI approach, and then combined with more specific application and algorithm descriptions. Taxonomy is from Reeder, at al (2022).[15]

Human-AI Interaction:

This field should contain a list of Human-AI touchpoints, interactions, and Interdependencies; places in the solution workflow where humans and/or AI work independently, are dependent on (or require inputs from) each other and/or where those interdependencies have to be effectively managed to achieve the system's overall objective. It should contain a bulleted list of key

[15] From Reeder, F., Kotras, D., Lockett, J., Pomales, C., & Lokas, R. (2022). *The Future State of Test and Evaluation of Artificial Intelligence - Enabled Systems in the Department of Defense.* MITRE Technical Report, MITRE Corporation, Mclean, Virginia.

human-AI interdependencies or interactions ranging from a single point of contact with the system (if the human simply receives the system output), or multiple points of contact (for example, if the system accepts feedback while performing the task, or has multiple potential interactions). Examples of types of human-AI interaction include independent (or no interaction), dependent (or relatively independent work that requires coordination and/or is contingent on other actors), and interdependent (collaborative work that should not or cannot be completed individually).

Expected Effects:

This field should contain a single-paragraph or list-based description of the expected outcomes, effects and/or impacts of using the system. It can include both low-level effects ("The classifier can classify incoming images at 95% accuracy") and high-level effects ("Fatalities are reduced"). It may also include negative effects if those are already anticipated ("Additional training will be needed for initial adoption to succeed").

Appendix C: Process for Querying the Use Case Repository and Identifying Similar Use Cases

The Use Case Repository[16] stores the collected Use Case Descriptions and their corresponding RDP-AIA Priority Report.

The repository is used in the DISCOVER phase, as part of its overall task to translate into a defined set of assurance needs the unstructured concerns uncovered in the IDENTIFY phase. In DISCOVER, the team retrieves cases from the Use Case Repository, drawing on the Use Case Description completed during pre-interview and UNDERSTAND (see Section 4.3.3) to find cases that are *similar* to the stakeholder case *on one or more critical dimensions* (see Figure 2 for an illustration). Being able to cross-reference the current case with previous similar cases allows the team to compare and contrast the AIA needs identified from past Use Cases that are similar and, where documented) to learn from how those concerns were previously addressed.

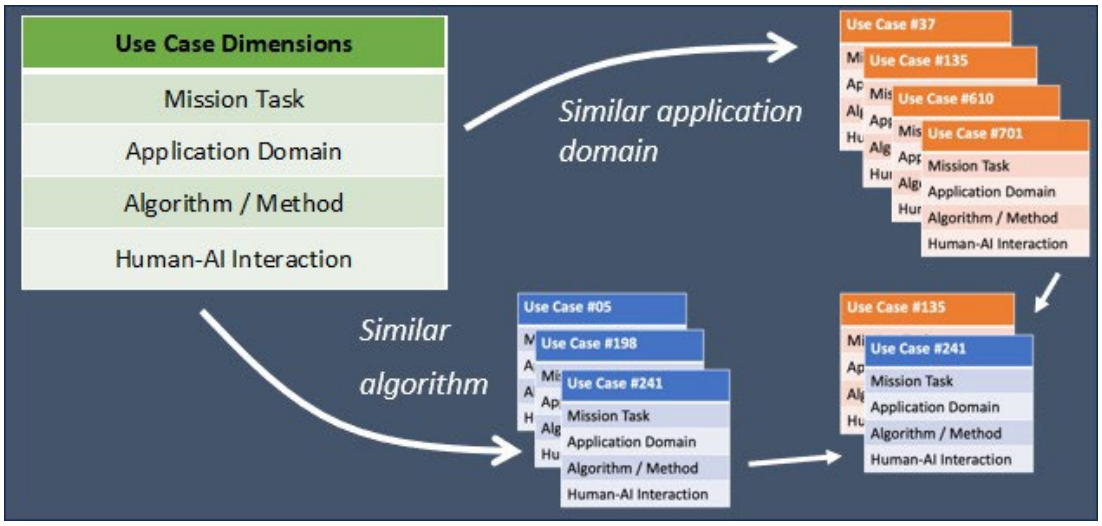


Figure 2: Retrieval of similar use cases. Initial retrieval may simply rely on common field attributes, such as similar application domains or algorithms, but over time more sophisticated retrieval techniques may be viable.

Finding similar cases can be done in several ways. Initially, while the population of use cases is small (<25), retrieval will mostly be either manual or use simple category-based similarities based on matching *AI Algorithm Types* (see Section 4.1.6), *AI Application Areas* (see Section 4.1.3), and/or mission/application *Sectors* (see Section 4.1.2). Over time, however, as the repository grows and the process for entering new cases becomes more standardized, more sophisticated measures of similarity, such as topic modeling, LLM-based retrieval, or other ML-based techniques may become viable.

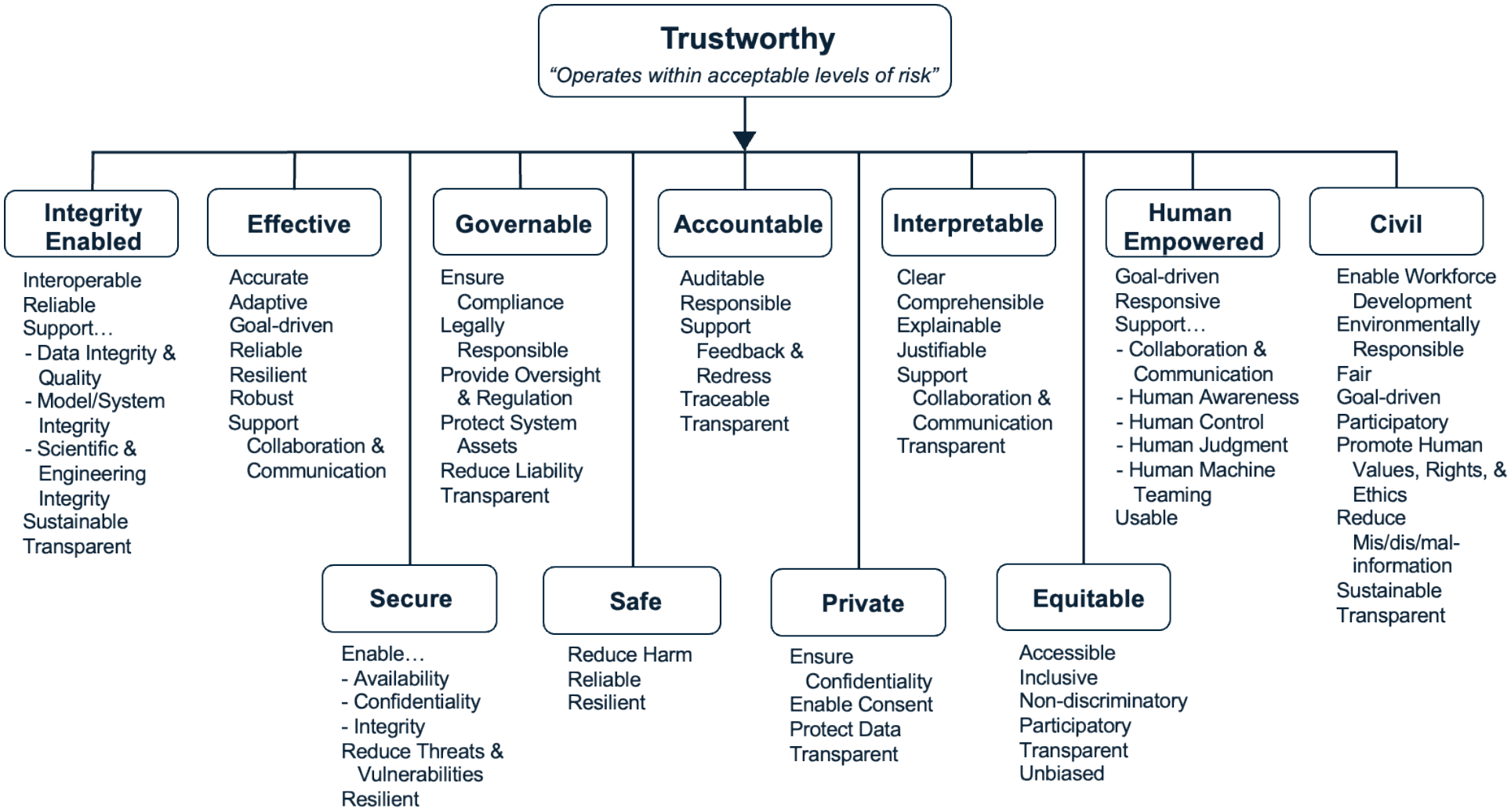
[16] The Use Case Repository is a future effort that will contain each of the Use Cases (incl. the associated RDP-AIA Priority Report and any related documentation) generated via the RDP-AIA. Use Cases in the Use Case Repository will be indexed by problem/solution features and assurance needs to permit identification of similar cases.

Four methods are proposed below. The first two methods are used in the version 1.1 of the RDP-AIA (i.e., this version). The other two are proposed as experimental retrieval methods, to be developed once the repository has reached >25 cases:

1. **Manual retrieval:** Retrieval by manual skimming of available use cases.
2. **Category-based retrieval:** Retrieval using the categories given in the Application Area, Algorithm Type, and/or Sector fields.
3. **Topic model or other NLP-based retrieval:** Retrieval by similar descriptions using either statistical NLP techniques, such as topic modeling and entity extraction.
4. **LLM-based retrieval:** Use an index vector of the entire case to perform retrieval of similar cases.
    - **LLM-based interaction**: Use an LLM-based system to interact with the case base in a conversational pattern.

Note that comparison with past similar cases is not meant simply to allow reuse of previous solutions for similar assurance issues—although, that may be very helpful when possible. Aligned *differences* between the stakeholder case and previous use cases may be just as useful. They may provide contrasts between cases that permit better isolation of issues that make the stakeholder case unique. They may also serve as a method of reframing the issues by analogy to a previous case, for example, in terms of the level of similarity of human-AI interaction pattern (e.g., AI in the loop, human in the loop, human on the loop, human out of the loop). Figure 2 illustrates how multiple dimensions could inform the identification of analogous cases.

Appendix D: AIA Landscape



Appendix E: AIA Landscape Glossary

Need Category	Specific Need	*Definition*
Integrity-Enabled		***Satisfies expectations for technical and scientific integrity.***
Integrity-Enabled	Support Collaboration & Communication	*Supports the ability of individual or group(s) of entities to work and/or exchange information with multiple other entities.* ***Qualifier:*** *Supports interaction across and interdependencies between various devices to perform common processes and achieve shared goals.*
Integrity-Enabled	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Produces repeatable processes and reproduceable outcomes.*
Integrity-Enabled	Support Data Integrity & Quality	*Supports the ability to assess and maintain completeness, consistency, accuracy, reliability, representativeness, and quality of data and data sources throughout its lifecycle, and in storage, during processing, and while in transit.*
Integrity-Enabled	Support Model/System Integrity	*Supports the ability to assess and maintain the soundness of a model or system's architecture, operations, and/or outcomes across its lifecycle, such that it performs as intended, unimpaired, and free from unauthorized manipulation.*
Integrity-Enabled	Support Scientific & Engineering Integrity	*Enables those who build and implement AI systems to be guided by established professional and scientific values and practices.*
Integrity-Enabled	Sustainable	*Processes are in place to ensure that the system can persist and be adapted over time to meet the needs of the communities in which it is deployed.* ***Qualifier****: Ensuring that data and system integrity are maintained over time.*
Integrity-Enabled	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use):* ***Qualifier:*** *Information needed to determine, test, and evaluate data, system/model, and scientific/engineering integrity is available as and when needed.*
Effective		***Achieves intent and/or desired outcomes.***

Effective	Accurate	*The capability to maintain closeness of results of observations, computations, or estimates to the true values or the values accepted as being true.*
Effective	Adaptive	*The capability to be responsive to change, including the ability to determine when current understanding, plans, or goals have deviated from expectations and/or the ability to achieve intent via alternative means.*
Effective	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers mission-relevant goals and is aligned with the organization's mission-relevant objectives in the context of risk tolerance levels and professional responsibility.*
Effective	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Consistently performs as expected.*
Effective	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards.* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that permits normal function to continue.*
Effective	Robust	*The capability of a system to maintain operations, performance, and/or expected impact under a variety of circumstances.*
Effective	Support Collaboration & Communication	*Supports the ability of individual or group(s) of entities to work and/or exchange information with multiple other entities.* ***Qualifier:*** *Supports interaction across and interdependencies between multiple internal and/or external entities.*
Secure		***Resistant to unauthorized activities***
Secure	Ensure Availability	*The capability to ensure timely and reliable access to and use of information.*
Secure	Ensure Confidentiality	*The capability to preserve authorized restrictions on information access and disclosure.* ***Qualifier:*** *Including means for protecting proprietary information.*
Secure	Ensure Integrity	*The capability to guard against improper information modification or destruction and ensure information non-repudiation and authenticity.*

Secure	Reduce Threats & Vulnerabilities	*Incorporates protocols to avoid, protect, respond, and recover from system weaknesses and both adversarial and non-adversarial threats.*
Secure	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards:* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that secures against and permits recovery from deliberate attacks, accidents, or naturally occurring threats or incidents.*
Governable		***Implements a framework of policies, rules, and processes for appropriate oversight within and across relevant organizations.***
Governable	Ensure Compliance	*Regulatory procedures are in place to prevent and address any divergence from standards and regulations.*
Governable	Legally Responsible	*Regulatory procedures are in place to identify individuals or entities at fault for harm caused by the system or other legal breaches.*
Governable	Provide Oversight & Regulation	*Regulatory procedures are in place to ensure that a diverse body of stakeholders identifies standards and regularly assesses system operations against them.*
Governable	Protect System Assets	*Regulatory procedures are in place to identify parties responsible for guarding and overseeing internal and external system (including third-party) assets and components.*
Governable	Reduce Liability	*The capability to assess potential failures to prepare and minimize the need for legal recourse and compensation, and permit insurability.*
Governable	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use). Qualifier: Information needed to oversee the system's operation, and for external parties to assess the oversight of the system, is available when needed.*
Safe		***Does not lead to a state in which human life, health, property, or the environment is endangered.***
Safe	Reduce Harm	*Built and tested to prevent misuse and avoid unintended harms of all types.*
Safe	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Consistently minimizes the potential for harm.*

Safe	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards.* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that maintains operations within acceptable levels of safety.*
Accountable		***Answerable to the stakeholders it empowers and to those it impacts for its proper and appropriate functioning, and obligated to address identified deficiencies.***
Accountable	Auditable	*The capability to periodically document, review, and evaluate the AI solution, assess its impacts, and provide on-demand access to information needed to determine the extent to which specified requirements are fulfilled.*
Accountable	Responsible	*Decisions about AI system development and use are aligned with intended aims and values, and recognize the unique influence they exert on people and society*
Accountable	Support Feedback & Redress	*Provide the opportunity for all Stakeholders, including individuals who are potentially impacted, to provide feedback, address concerns, and engage in procedures designed to change aspects of the system in ways that improve, rectify, repair, and/or remedy impacts (e.g., reporting problems, appealing system outcomes, and opt out of system processes).*
Accountable	Traceable	*Processes and outcomes can be monitored and traced back to simple root causes, or in complex situations, traced to potentially multiple and non-linear causes.*
Accountable	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier:*** *Stakeholders including impacted communities have appropriate access to information about values, choices, and intentions behind the system.*
Private		***Safeguards information collection and use to preserve autonomy and dignity.***
Private	Enable Confidentiality	*The capability to restrict data access to protect personal privacy and proprietary information.* ***Qualifier:*** *Including means for protecting personal privacy.*
Private	Enable Consent	*Individuals must explicitly agree to the processing of personally relevant data and be informed of risks and options.*
Private	Protect Data	*All parts of the system lifecycle are designed to protect the rights of data subjects.*

Private	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier:*** *Stakeholders are made aware of data processing practices and associated risks, and any personally relevant information processed by the system.*
Interpretable		***Makes processes and outputs apparent and meaningful in the context of functional and anticipated purposes.***
Interpretable	Clear	*The system presents its processes and outputs such that the human can readily incorporate them into the workflow.*
Interpretable	Comprehensible	*The capability to provide users with access on-request to sufficient contextual information (e.g., system goals, objectives, inputs, assumptions, expected operating conditions, constraints) to allow them to develop a meaningful and up-to-date mental model of the system (i.e., integrate situational context with their own knowledge, understanding, goals, values, and preferences), in a way that permits evaluation of the appropriateness of system operations and outputs and/or anticipation of its behavior.*
Interpretable	Explainable	*The capability to provide a description of how or why an output was produced that captures the reasoning process(es) and/or technical mechanism(s) that actually led to the outcome, along with supporting evidence.*
Interpretable	Justifiable	*The capability to provide an adequate reason (e.g., moral rationale) for producing a particular outcome that is capable of withstanding scrutiny, without necessarily providing a causal explanation.*
Interpretable	Support Collaboration & Communication	*Supports the ability of individual or group(s) of entities to work and/or exchange information with multiple other entities.* ***Qualifier:*** *Supports building common ground vertically (across echelons) and horizontally (across units) to permit understanding of the 'bigger picture.'*
	Transparent	*The capability to make explicit the functions, operations, and outcomes (incl. data, algorithms, and models in use). Qualifier: Stakeholders have appropriate access to required information about the AI system's processes and outputs.*
Equitable		***Addresses disparities in use and outcomes across individuals and groups.***

Equitable	Accessible	*Supports comparable ease of use and access across all users.*
Equitable	Inclusive	*Processes and methods are included that consider the demographic diversity and diverse user experiences of those communities for whom the system is designed.*
Equitable	Non-discriminatory	*Processes are in place to ensure that individuals and groups with similar non-protected characteristics are assigned similar outputs; differences in protected characteristics should not cause significant differences in outputs.*
Equitable	Participatory	*Processes are in place to support engaging, across the entire AI lifecycle, with Stakeholders that represent a broad range of perspectives, including those from potentially impacted communities.* ***Qualifier:*** *Ensure marginalized communities are included to reduce inequity.*
Equitable	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier****: Any discrepancies in treatment among individuals and groups are clearly communicated.*
Equitable	Unbiased	*Any systematic preference for or against some group of impacted people due to data or models is identified and mitigated as much as possible.*
Human-Empowered		***Leverages human capabilities and enables pursuit of human goals.***
Human-Empowered	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers the operator's goals in the context of broader operational, strategic, and societal goals.*
Human-Empowered	Responsive	*The capability to promptly probe and obtain answers from and about the AI system, including its development, intentions, operations, outputs, and associated explanations.*
Human-Empowered	Support Collaboration & Communication	*Supports the ability of individual or group(s) of entities to work and/or exchange information with multiple other entities.* ***Qualifier:*** *Facilitates shared understanding and workflows among diverse stakeholders.*
Human-Empowered	Support Human Awareness	*Humans know when they are interacting with or are affected by AI, and know which tasks an AI is performing where they are out of the loop.*

Human-Empowered	Support Human Control	*Humans can direct resources, activities, and priorities as needed and, where necessary, can modify or take over an AI's decisions or actions.*
Human-Empowered	Support Human Judgment	*Humans are engaged in an AI's decision process(es) throughout the AI lifecycle, and especially during operations.*
Human-Empowered	Support Human Machine Teaming	*Adaptive, bi-directional team interaction among humans and machines that augments human capabilities for improved mission outcomes.*
Human-Empowered	Usable	*User interfaces are easy to use, efficient, memorable, learnable, and minimize and permit recovery from error, and are considered satisfactory by those who need to interact with them.*
Civil		***Designed and operates in accordance with social norms and the public good.***
Civil	Enable Workforce Development	*Supports human jobs, economies, and AI workers, and their development, without putting them at risk.*
Civil	Environmentally Responsible	*Actively protects or, at least, does not represent a threat to the environment and/or broader ecosystem.*
Civil	Fair	*The system benefits society as a whole and does not contribute to or perpetuate social imbalances.*
Civil	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers the goals of communities in which the system is deployed.*
Civil	Participatory	*Processes are in place to support engaging with Stakeholders across the entire AI lifecycle that represent a broad range of perspectives, including those from potentially impacted communities.* ***Qualifier:*** *Impacted communities play a key role in developing and sustaining the system.*
Civil	Promote Human Values, Rights, & Ethics	*The system works in humanity's best interests; supports, observes, and does not conflict with commonly held human values, ethics, rights, and societal norms.*
Civil	Reduce Mis/dis/mal-information	*The capability to manage context and content to reduce risk of manipulation and polarization of opinions and beliefs.*

Civil	Sustainable	*Processes are implemented to ensure that the system can persist and be adapted over time to meet the needs of the communities in which it is deployed.* ***Qualifier:*** *Ensures that the system continues to be accepted over time by the communities in which it is deployed.*
Civil	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier****: Documents and communicates to respective parties expected and actual impacts on communities.*

Appendix F: Sector Explanation

This appendix describes how the sectors used in UNDERSTAND Step 2 were derived from the UN Classification of the functions of government (COFOG), first developed in 1991 and last updated and simplified in 2019. The COFOG comprises 11 divisions with multiple groups in each division and multiple classes in each group. We adapted the COFOG to emphasize domains frequently encountered in our work. This manner of adaptation is generally in accordance with the COFOG's recommended use. We are not experts on the COFOG and have adapted it based on our own introductory knowledge of the framework. Primarily we made three kinds of adaptations:

- Renamed a division to emphasize certain groups within that division
- Extracted one or more groups from a division to a separate category
- Merged two or more groups and/or divisions

Category	How derived	Example sponsor / U.S. Gov orgs
1. National security	Approximately equivalent to Division 02 Defense.	DOD
2. Intelligence and information operations	Extracted as a piece from Group 02.1 Military Defense to emphasize a frequent sponsor work domain	USGC, NSA
3. Law enforcement, judicial, prisons	Approximately equivalent to Division 03 Public Order and Safety	DHS, DOJ
4. Emergency management	Extracted and combined various groups including 02.2 Civil Defense, 03.2 Fire Protection, 07.3 Hospital Services, and others to emphasize a sponsor domain of concern	DHS
5. Health and wellness	Approximately equivalent to Division 07 Health	VHA, DHA, CMS
6. Administration and finance	Approximately equivalent to Division 01 General Public Services	IRS, VBA
7. Transportation	Extracted Group 04.5 Transport	FAA, DOT
8. Industrial and manufacturing	Extracted Group 0.4.4 Mining, Manufacturing, and Construction	DOC, NIST
9. Environment, energy, agriculture	Extracted Group 04.2 Agriculture, Forestry, Fishing and Hunting; Group 04.3 Fuel and Energy; merged with Division 05 Environmental Protection	DOE

10. Communications	Extracted Group 04.6 Communication	FCC
11. Space	Added (not called out in COFOG) for cases primarily targeting space exploration and development; briefly mentioned in Class 04.5.4 Air Transport	NASA, NOAA
12. Education, recreation, culture	Merged Division 08 Recreation, Culture, and Religion with Division 09 Education	DOEd
13. Housing and community amenities	Equivalent to Division 06	HHS
14. Social protections (equity, unemployment)	Equivalent to Division 10	GSA, DOL
15. Other (Write in)		
16. Domain-agnostic		

Appendix G Interview Aids

G.1 Working with AIA Needs

Use in Phases (Steps): IDENTIFY (All Steps), DISCOVER (All Steps), and PRIORITIZE (Step 1).

1. As the AIAD Lab team identifies Stakeholder AIA needs in the IDENTIFY phase, add them to the table below, filling in the particular assurance need (in Stakeholder terms), their rationale, and any associated notes in the respective columns.
2. As the Stakeholder's needs are translated into specific needs from the AIA landscape in the DISCOVER phase, fill in the AIA Needs Category and Specific Needs columns. Some initial concerns or needs may map to more than one category. There may also be new needs not derived from AIA landscape in the IDENTIFY phase. Add new rows to the table and duplicate content as needed.
3. Before prioritizing needs in PRIORITIZE Step 1, select all columns except Needs Category and Specific Needs, go to the Home tab in Word, click the corner arrow in the Font section, and select the "hidden" effect. This will avoid discussions over exact wording when the Stakeholder sees the table.
4. Holding Ctrl+Shift, use the arrow keys to reorder the rows as needed.
5. Delete all but the top ~6 concerns (or a different number as agreed with the Stakeholder).
6. After finishing PRIORITIZE Step 1, un-hide the columns and paste the result into the RDP-AIA Priority Report Template.

AI Assurance Needs		Assurance Need ("What")	Rationale ("Why")	Notes
Category	Specific Need			
e.g., Interpretable	*e.g., Explainable*	*e.g., "User needs to understand the reasoning behind recommendations."*	*e.g., "Previous systems failed because users could not determine reasons for recommendations"*	

G.2 Prioritizing AIA Needs by Risk

Use in Phase (Steps): PRIORITIZE (Steps 4 and 5).

1. As the Stakeholder assigns Likelihood and Severity to each worst credible impact (see Appendix H), copy the appropriate ratings label from the colored labels (Very Low, Very Low to Low, Low, Low to Moderate, etc.) and paste into the appropriate cell in the table.
2. Derive the risk level from the risk matrix below and copy the appropriate colored label into the Risk column.
3. As the Stakeholder states their Risk Tolerance level, copy the appropriate label from the colored labels (Very Low, Very Low to Low, Low, Low to Moderate, etc.) and paste into the appropriate cell in the table.
4. Holding Ctrl+Shift, use the arrow keys to reorder the rows as needed.
5. Where necessary, repeat for PRIORITIZE Step 5 (Risk "At Deployment"). After each step, copy and paste into the appropriate location in the RDP-AIA Priority Report (Risk "As-Is" vs. Risk "At Deployment").

Very Low *Very Low to Low* *Low* *Low to Moderate* *Moderate* *Moderate to High* *High* *High to Very High* *Very High*

AI Assurance Need		Worst Credible Impact		Risk	Risk Tolerance	Notes
Category	Specific Need	Likelihood	Severity			
e.g. Interpretable	*e.g., Explainable*	*Low*	*Very High*	*Moderate*	*Low*	

Severity

Likelihood	Very Low	Low	Moderate	High	Very High
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low

G.3 Risk Assessment Cards

The Risk Assessment Cards can be a useful alternative for prioritizing needs by risk when conducting interviews in-person or using a virtual collaboration tool that allows the Stakeholder to arrange items spatially, like Mural.

Use in Phase (Steps): PRIORITIZE (Steps 4 and 5).

1. As the Stakeholder assigns Likelihood and Severity to each worst credible impact, fill in the appropriate cells in the template.
2. Derive the risk level from the risk matrix below and copy the appropriate colored label into the Risk row.
3. In PRIORITIZE Step 4, allow the Stakeholder to arrange the cards to reflect relative priorities and assess risk tolerance levels.
4. Repeat for PRIORITIZE Step 5 (Risk "At Deployment"). After each step, copy and paste into the appropriate location in the RDP-AIA Priority Report (Risk "As-Is" vs. Risk "At Deployment").

Assurance Category	[Assurance needs category]
Assurance Need	[Assurance specific need]
Impact	[Worst credible negative impact]
Severity	[Magnitude]: [Justification]
Likelihood	[Magnitude]: [Justification]
Risk	[Magnitude]
Acceptable Risk	[Magnitude]
Other Notes	[Free text]

Appendix H: Risk Assessment

This appendix supplements PRIORITIZE Steps 2 and 4.

H.1 Assessing "As-Is" and "At Deployment" Risk

The purpose of the risk assessment component of the PRIORITIZE phase of the RDP-AIA is to assess the risk associated with not assuring each of the ~6 specific identified needs (potentially rank-ordered IF the Stakeholder had 10 or more needs – see PRIORTIZE Step 1). This is achieved through Stakeholder judgments of (a) the magnitude of impact of the worst credible threat (i.e., severity) associated with not assuring a particular assurance need, and (b) the expected frequency of that impact (likelihood).

The following definitions and rating scales are used in the respective judgements:

Severity is defined as: *The magnitude of negative impacts on assets, operations, individuals, organizations, the Nation, or society more broadly* [NIST 800-30, Table H-3.] It is measured on the following scale.

Qualitative Values	Description of Severity
Very High	Multiple severe or catastrophic negative impacts
High	Severe or catastrophic negative impacts (major damage, loss, or harm)
Moderate	Serious negative impacts (significant damage, loss, or harm)
Low	Limited negative impacts (minor damage, loss, or harm)
Very Low	Negligible negative impacts

Likelihood is defined as: *The expected frequency of negative impacts, assuming the envisioned scale, frequency, and duration of solution use.* [NIST 800-30, Table G-3.]. Likelihood is measured on one of *two* scales (see Scale 1 & 2 below)—chosen by the participant. The same scale should be *used for all assurance needs being rated*.[17]

Scale 1: Likelihood (expressed in frequency of negative impacts per year):

Qualitative Values	Description of Likelihood (per year)
Very High	More than 100 times per year
High	Between 10-100 times per year
Moderate	Between 1-10 times per year
Low	Less than once per year but more than once every 10 years
Very Low	Less than once every 10 years

Scale 2: Likelihood (expressed in frequency of negative impacts per solution use):

Qualitative Values	Description of Likelihood (per use)
Very High	Once per use
High	Between once per use and once every 10 uses
Moderate	Between once every 10 uses and once every 100 uses
Low	Between once every 100 uses and once every 1,000 uses
Very Low	Less than once every 1,000 uses

[17] The prioritization of assurance needs—across Stakeholders who use a different likelihood scale will be similar (because the relative risks that result from use of each scale will be similar) but may not be identical. The absolute magnitudes of risk are NOT directly comparable between different participants/solutions that use different likelihood scales. In addition, a participant's acceptable levels of risk may differ depending on the likelihood scale being used.

Risk is defined as: *The potential for negative impact. Risk is assessed by a combination of Severity and Likelihood of negative impacts*. The following risk matrix (from on NIST's Risk Management Framework, NIST 800-30, Table I-2.) is used to derive an assessment of risk from ratings of Severity and Likelihood.

Severity

Likelihood	**Very Low**	**Low**	**Moderate**	**High**	**Very High**
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low

Note that Severity and Likelihood judgments are first made for the "As-is" solution (see Section 4.4.2), which may differ from judgments about risks associated with solution deployment. Once, "As-is" risk assessments have been generated and assurance concerns prioritized, Severity and Likelihood judgments are then made for projected risks "At Deployment" where applicable (see Section 4.4.4). Risk assessment cards discussed in Appendix G can be used to capture relative ratings of Severity, Likelihood, and Risk, along with other related information.

Appendix I: RDP-AIA Priority Report Template

A full Priority Report template with guidance on how to complete each part of the report is available on request as a separate document. The following example highlights the main information and tables in that full template. Instructions [*italicized, purple text, like this*] for what should be entered into each field, or examples of entries, are provided below.

Bottom Line Up Front:

Authors:

e.g., Paul Ward (wardp@mitre.org)		

Part 1: Project Summary

Project Information

Project/task name:		Project sponsor:		MITRE project/task lead:	

Use Case Description

Metadata	First created:	[*MM/DD/YYYY*]	Last updated:	[*MM/DD/YYYY*]	Sources:	Interviews with [*Insert name of POC/Personnel involved*]
Problem Description	Mission Problem:	[*Provide a brief description of the Stakeholder's Mission Problem that the AI Capability will help solve*]				
	Sector:	[*Select from the list provided; add further details]*		Application Area:	[*Select from the list provided; add further details]*	
Solution Description	Maturity:	[*Select from the list provided; add further details]*		Algorithm Type	[*Select from the list provided; add further details]*	
	Solution Workflow:	[*Add ordered list of 3-6 steps in the workflow, identifying key AI components, e.g. step 1, step 2, etc.*]				
	Human Interaction:	[*Add 3-6 points of human interaction; identify where in the workflow (e.g., between steps 1&2)*]		Expected Effects:	[*List of up to 6 expected effects on the mission*]	
AIA Analogous Cases:		[Where identified, add case #s for similar cases identified in the Use Case Repository; specify similar characteristics between cases]				

Hint
The following tables should be completed for the AI solution in its current stage of development (i.e., "As-Is"). IF, during the RDP-AIA, the Stakeholder indicated that the assurance needs would differ once deployed, THEN the information collected about those differences should be reported in a second set of tables/figures labeled: "FUTURE RISK ASSESSMENT ("AT DEPLOYMENT").

CURRENT RISK ASSESSMENT (SYSTEM "AS-IS")

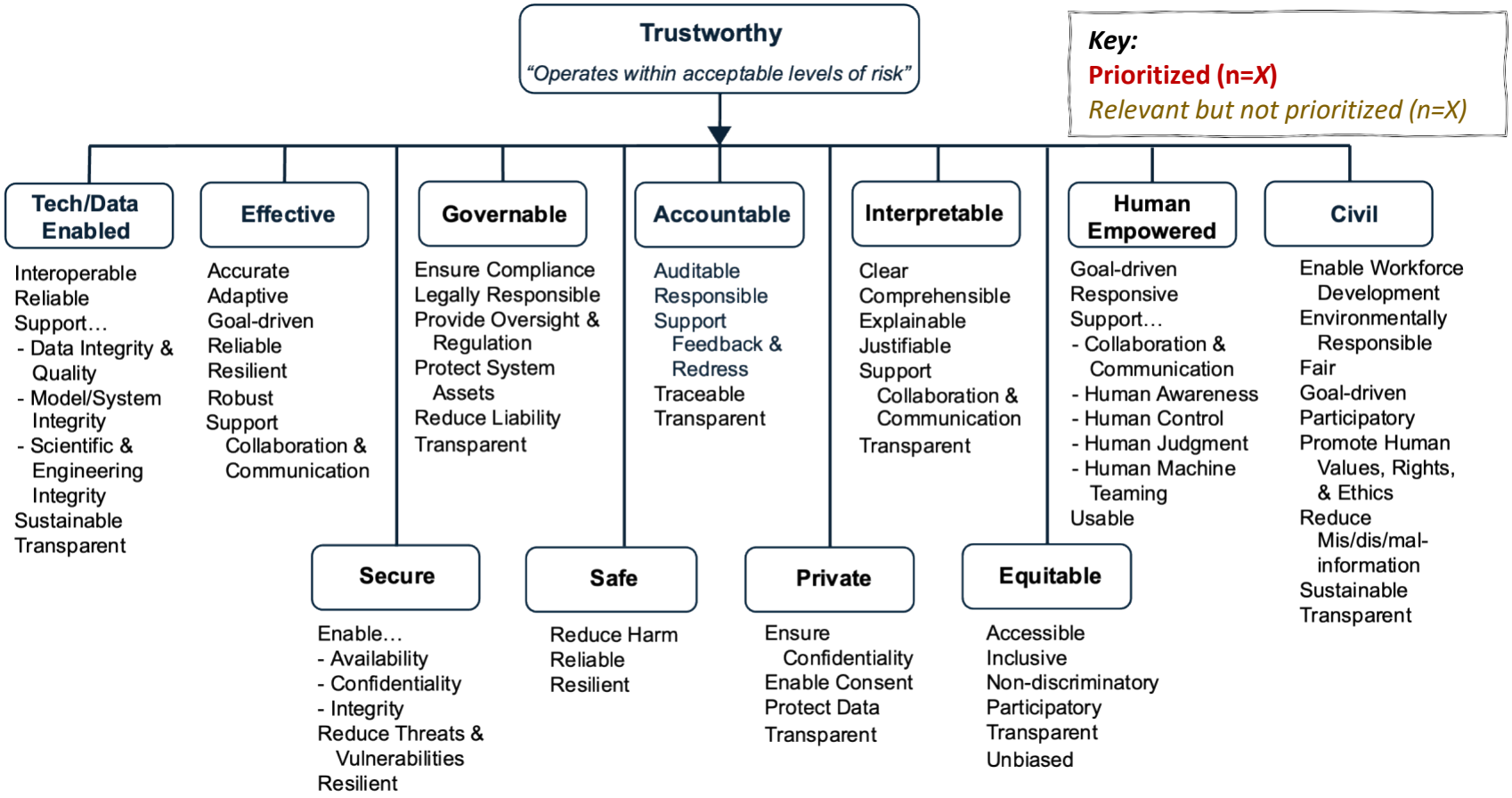
Part 2: AI Assurance Needs

Table 1. Priority AIA Needs (in alphabetical order by *Category*, then *Specific Need*)

AI Assurance Needs		Definition of Specific Need	Interview Context[18]
Category	Specific Need		
e.g., **Accountable**	e.g., Traceable	e.g. Processes and outcomes can be monitored and traced back to simple root causes, or in complex situations, traced to potentially multiple and non-linear causes.	[*Paraphrase relevant content from interview*]

[18] The *Interview Context* column includes a paraphrased excerpt from the Stakeholder interviews, which provides some initial insight into the rationale for each need being identified.

Figure 1. MITRE's AI Assurance Landscape[19]. Priority (red/bold font) and relevant (gold/italic font) assurance needs are highlighted.



[19] The MITRE AI Assurance Landscape contains 82 assurance needs in total (incl. 11 categories and 71 specific needs). A glossary of all terms is available upon request.

Part 3: Worst Credible Impacts

Table 2: Worst Credible Impacts Associated with Each Specific Need (in alphabetical order by *Category*, then *Specific Need*)

AI Assurance Needs		Worst Credible Impact
Category	Specific Need	
e.g., **Accountable**	e.g., Traceable	IF A, AND B, THEN X, AND Y

Part 4: Likelihood & Severity of Worst Credible Impacts

Table 3: Stakeholder's Severity and Likelihood Ratings of Worst Credible Impacts (in alphabetical order by *Category*, then *Specific Need*).

AI Assurance Needs		Worst Credible Impact	Likelihood	Severity	Additional Notes
Category	Specific Need				
e.g., **Accountable**	e.g., Traceable	IF A, AND B, THEN X, AND Y	Very Low	Very Low to Low	
			Very Low to Low	Low	
			Low to Moderate	Low to Moderate	
			Moderate	Moderate	
			Moderate to High	High	
			High to Very High	Very High	

Part 5: Assessed Risk and Risk Tolerance

Table 4: AIA Needs Prioritized by Risk and Risk Tolerance Levels (in order of priority) .

Priority	AI Assurance Needs		Risk	Risk Tolerance	Risk Relative to Risk Tolerance	Notes about Risk Tolerance
	Category	Specific Need				
1	Interpretable	Explainable	Low to Moderate (4)	Very Low to Low (2)	Risk **ABOVE** Acceptable Level (+2)	
2			Low to Moderate (4)	Low to Moderate (4)	Risk **AT** Acceptable Level (0)	
3			Low to Moderate (4)	Moderate (5)	Risk **BELOW** Acceptable Level (-1)	
3			Low to Moderate (4)	Moderate (5)	Risk **BELOW** Acceptable Level (-1)	
3			Low (3)	Low to Moderate (4)	Risk **BELOW** Acceptable Level (-1)	
3			Very Low (1)	Moderate (5)	Risk **BELOW** Acceptable Level (-4)	

FUTURE RISK ASSESSMENT (SYSTEM "AT DEPLOYMENT")

[*Repeat Tables 1-4 and Figure 1, IF assurance needs "at deployment" differ from those "as-is"*]

Part 6: Recommendations

Lessons Learned:

Appendix J: Example Emails

J.1 Scheduling email

Dear [Stakeholder],

Thank you for reaching out to the AI Assurance & Discovery (AIAD) Lab about [INSERT PROJECT NAME]. The first step in our process is to understand your mission problem and associated AI assurance needs. We have a structured process for eliciting this information: the Risk Discovery Protocol for AI Assurance (RDP-AIA). This protocol employs an interview-based data and information collection method. It is designed to be implemented in two 2-3 hour sessions, one on each of two separate days (i.e., ~6 hours in total) depending on the complexity of your use case and extent of assurance concerns .

To schedule the interview, please let us know if any of the times listed below work for you or if you have a preferred time/date not listed. Feel free to invite others who can speak to the AI assurance concerns on your project.

- *[Timeslot 1]*
- *[Timeslot 2]*
- *[Timeslot 3]*

After scheduling the interview, we will send you a short list of questions about your project, which should take between 15-30 minutes to answer. Please allow yourself enough time to respond to these questions before attending the interview. We would appreciate it if you could send your responses to these questions at least two days prior to the scheduled date.

We look forward to speaking with you about your AI assurance needs.

Respectfully,

The AIAD Lab team

J.2 Pre-Interview Questions Email

Hello [Stakeholder],

We look forward to speaking with you at [date and time].

To help provide context for our discussion, we would be grateful if you could share relevant background materials (e.g., ppt deck, white paper) with the AIAD Lab team before [INTERVIEW TIME/DATE] to provide an overview of your specific mission problem and the associated AI solution. Please send these to us as an email attachment.

Please answer the following questions to the best of your ability:

Please choose a response to each question from the options provided. Please send your responses at least TWO days prior to the scheduled interview date. *Where multiple choices are given, pick the best fit, or select "other" and write your response. In any case, additional description is welcome but not required.*

1. *What is your mission problem or specific challenge your system addresses (i.e., what is your solution trying to achieve)? Please answer as a brief paragraph (5 or fewer sentences).*

2. *In which sector or domain will your AI solution be deployed?*
    i. *National security*
    ii. *Intelligence and information operations*
    iii. *Law enforcement, judicial, prisons*
    iv. *Emergency management*
    v. *Health and wellness*
    vi. *Administration and finance*
    vii. *Transportation*
    viii. *Industrial and manufacturing*
    ix. *Environment, energy, agriculture*
    x. *Communications*
    xi. *Space*
    xii. *Education, recreation, culture*
    xiii. *Housing and community amenities*
    xiv. *Social protections (equity, unemployment)*
    xv. *Other [please State]*
    xvi. *Domain-agnostic*

3. *What is the Application Area for the AI-enabled system you are developing?*
    i. *Computer Vision*
    ii. *Natural Language Processing*
    iii. *Planning & Scheduling*
    iv. *Robotics & Autonomous Systems*
    v. *Decision Aids & Predictive Analysis*
    vi. *Other [please State]*

4. *In which stage of development (e.g., Technology Readiness Level) is your AI solution?*
    i. *Early Development*
    ii. *Late Development*
    iii. *Deployed*
    iv. *Other [please State]*

5. *Which type of AI method(s) or algorithm(s) will(/have) you use(d) in your AI solution? (Multiple choices permitted.)*
    i. *Knowledge Representation & Reasoning*
    ii. *Machine Learning – Supervised*
    iii. *Machine Learning – Unsupervised*
    iv. *Machine Learning – Semi-supervised*
    v. *Machine Learning – Reinforcement Learning*
    vi. *Other [please State]*

Many thanks for your input.

Best regards,
The AIAD Lab team

J.3 Follow-up Email

Hello [Stakeholder],

Thank you for discussing your AI assurance needs with us on [date]. We'd appreciate your final review before submitting our report, which serves as input to the next step of the AIAD Lab process: Measuring and mitigating your assurance needs. There are four items below for your review.

1. *First, please check our description of your use case for any inaccuracies, glaring omissions, or items that might have changed since our interview session. You may suggest any changes you want, but our goal is to be concise and minimize errors.*

[AIAD Lab Team: Paste Use Case Description from the RDP-AIA Priority Report in this table.]

Problem Description	Mission Problem:	[*Provide a brief description of the Stakeholder's Mission Problem that the AI Capability will help solve*]		
	Sector:	[*Select from the list provided; add further details]*	Application Area:	[*Select from the list provided; add further details]*
Solution Description	Maturity:	[*Select from the list provided; add further details]*	Algorithm Type	[*Select from the list provided; add further details]*
	Solution Workflow:	[*Add ordered list of 3-6 steps in the workflow, identifying key AI components, e.g. step 1, step 2, etc.*]		
	Human Interaction:	[*Add 3-6 points of human interaction; identify where in the workflow (e.g., between steps 1&2)*]	Expected Effects:	[*List of up to 6 expected effects on the mission*]

2. *Next, please review the following two lists of AIA Needs--As-Is and At-Deployment--that we generated during our interview and confirm that these still make sense to you and no further*

changes are needed. If helpful, you can browse the attached AI Assurance Glossary for the definitions. While browsing, let us know if any items stand out to you that are not captured here.

[AIAD Lab Team: Attach the Glossary to the email.]

[AIAD Lab Team: Paste Tables 1-4 for As-Is and At Deployment (where applicable) here]

3. *Finally, please review the recommendations and lessons learned that you shared with us and let us know if you have anything further to add.*

[AIAD Lab Team: Paste Recommendations and Lessons Learned from the RDP-AIA Priority Report here.]

4. *Our report will be added to a repository and could be cited in discussions with other AIAD Lab customers who might be addressing similar problems.* ***Do you see anything above that you would prefer NOT be shared with other parties?***

Thanks again for your time interviewing and your follow-up with these items. As soon as we receive your response, we will submit the report so you can work with the AIAD Lab to measure and mitigate these AI assurance risks.

Best regards,
The AIAD Lab team