

The logo features the word "RESPONSIBLE AI" in a blue, sans-serif font. The letter "R" is significantly larger and stylized, with a lighthouse icon on top and three signal waves extending to the right. A vertical line separates "RESPONSIBLE AI" from "ToolKit". "ToolKit" is in a bold, black, sans-serif font, with a blue toolbox icon containing a wrench and a screwdriver to its right.

RESPONSIBLE AI | **ToolKit**

GENERATIVE AI VERSION
1.0



CDAO

Accelerate DOD's adoption
of data, analytics, and AI to
generate decision advantage.

Responsible AI Division

Contents

- INTRODUCTION AND INSTRUCTIONS 6**
- OVERVIEW OF RAI ACTIVITIES THROUGHOUT THE PRODUCT LIFE CYCLE..... 7**
- STAGE 1. INTAKE 9**
 - 1.1 CONSIDER PREVIOUSLY LEARNED LESSONS..... 9
 - 1.2 DETERMINE RELEVANT LAWS, ETHICAL FRAMEWORKS, AND POLICIES 9
 - 1.3 IDENTIFY AND ENGAGE STAKEHOLDERS 9
 - 1.4 CONCRETIZE THE USE CASE FOR THE AI10
 - 1.5 DECIDE TO PROCEED TO IDEATION.....10
- STAGE 2. IDEATION.....11**
 - 2.1 DEFINE REQUIREMENTS11
 - 2.2 IDENTIFY RISKS & OPPORTUNITIES / NAVIGATE TRADEOFFS11
 - 2.3 WRITE STATEMENTS OF CONCERN.....11
 - 2.4 DESIGN TO REDUCE ETHICAL/RISK BURDENS.....12
 - 2.5 ACCOUNTABILITY, RESPONSIBILITY, & ACCESS FLOWS AND GOVERNANCE12
- STAGE 3. ASSESSMENT 14**
 - 3.1 ASSESS REQUIREMENTS, STATEMENTS OF CONCERN, MITIGATIONS, AND METRICS14
 - 3.2 EXPLORATORY DATA ANALYSIS (EDA)15
 - 3.3 UPDATE AI SUITABILITY, FEASIBILITY, AND ADVISABILITY ASSESSMENTS16
 - 3.4 UPDATE DOCUMENTATION.....16
- STAGE 4. DEVELOPMENT/ACQUISITION 17**
 - 4.1 INSTRUMENT AI TO PROMOTE ASSURANCE.....17
 - 4.2 UPDATE DOCUMENTATION.....19
- STAGE 5. TEVV 20**
 - 5.1 TEST SYSTEM FOR ROBUSTNESS, RESILIENCE, AND RELIABILITY.....20
 - 5.2 UPDATE DOCUMENTATION21
- STAGE 6. INTEGRATION & DEPLOYMENT..... 22**
 - 6.1 PERFORM OPERATIONAL TESTING22
 - 6.2 TRAIN USERS22
 - 6.3 ESTABLISH INCIDENT RESPONSE PROCEDURES22
 - 6.4 AUDITING & OVERSIGHT MECHANISMS22
 - 6.5 UPDATE DOCUMENTATION23
- STAGE 7. USE 24**
 - 7.1 PERFORM CONTINUOUS MONITORING OF THE SYSTEM AND ITS USE, CONTEXT, AND ECOSYSTEM.....24
 - 7.2 ENSURE UPDATING AND RETRAINING24
 - 7.3 PLAN FOR SYSTEM RETIREMENT25
 - 7.4 RECORD LESSONS LEARNED.....25
- THE RESPONSIBLE AI TOOLS LIST26**
 - RAI USE CASE REPOSITORY27
 - AI INCIDENT REPOSITORY27
 - AI INCIDENTS DATABASE27
 - FAILURE MODES RESOURCES.....27

DoD AI GUIDE ON RISK (DAGR).....	27
IBM DESIGN THINKING TOOLKIT.....	28
DESIGN KIT INSPIRATION METHODS.....	28
18F DISCOVER METHODS.....	28
STAKEHOLDER MAPPING TEMPLATE.....	28
USE CASE ANALYSIS.....	28
IBM DATA PRIVACY TOOLKIT.....	29
CONCEPT OF OPERATIONS EXAMPLE.....	30
AETHER DATASHEET TEMPLATE.....	30
MODEL CARD EXAMPLES.....	31
HUGGING FACE DATA CARD TEMPLATE.....	31
HUGGING FACE MODEL CARD TEMPLATE.....	32
PUBLIC AFFAIRS AND COMMUNICATIONS TOOLKIT.....	32
RESPONSIBILITY FLOWS TOOL.....	32
PLANNING WORKSHEET FOR DIU RAI GUIDELINES.....	33
ALGORITHMIC IMPACT ASSESSMENT.....	33
FRAMEWORK FOR ETHICAL DECISION MAKING.....	33
IBM FAIRNESS 360.....	34
FAIRML.....	36
WHAT-IF TOOL.....	37
WORD EMBEDDING ASSOCIATION TASKS (WEATS) METHOD.....	38
BIAS AND FAIRNESS AUDIT TOOL.....	38
HUMAN BIAS RED-TEAMING TOOLKIT.....	38
COGNITIVE BIASES RESOURCE.....	39
FEATURE RELEASE ROLLBACK RESOURCE.....	39
RAI ACQUISITION TOOLKIT.....	39
RAI PROGRAM MANAGER REVIEW.....	39
IDA HUMAN-MACHINE TEAMING GUIDEBOOK.....	39
MIT-LL HMT RED-TEAMING GUIDEBOOK.....	40
HUMAN-MACHINE TEAMING SYSTEMS ENGINEERING GUIDE.....	40
TRUST IN AUTONOMOUS SYSTEMS TEST.....	40
XAI TOOLKIT - SALIENCY.....	41
LIME 42	
SHAP 43	
EXPLAINERDASHBOARD.....	44
DIVERSE COUNTERFACTUAL EXPLANATIONS (DICE).....	45
SHAPASH.....	45
MODEL AGNOSTIC LANGUAGE FOR EXPLORATION AND EXPLANATION (DALEX).....	46
INTERPRETML.....	47
IBM EXPLAINABILITY 360.....	48
PYTHON OUTLIER DETECTION (PYOD).....	49
ALICE50	
EQUI(NE2).....	50
IBM UNCERTAINTY QUANTIFICATION 360.....	50
ADVERSARIAL PATCHES REARRANGED IN CONTEXT (APRICOT).....	51
ARMORY TESTBED.....	51
ALIBI DETECT.....	52
BIAS BOUNTY GUIDEBOOK.....	52
EXECUTIVE DASHBOARD.....	52
INCIDENT RESPONSE GUIDANCE.....	52
INDIELABEL END-USER AUDIT.....	53
THREAT MODELING RESOURCE.....	53
ROOT CAUSE ANALYSIS.....	53

STANFORD WILDS DATASET	54
DOMAIN GENERALIZATION DATASET LIST.....	55
CHECKLIST FOR ML SUITABILITY	55
GAMECHANGER POLICY DATABASE.....	55
CARBON COSTS CALCULATOR	55
TENSORFLOW FAIRNESS INDICATORS	56
TENSORFLOW MODEL REMEDIATION	56
FOOLBOX.....	57
COUNTERFIT	57
SMARTNOISE	57
IBM ADVERSARIAL ROBUSTNESS 360 ATTACKS	58
IBM ADVERSARIAL ROBUSTNESS 360 DEFENSES.....	59
IBM ADVERSARIAL ROBUSTNESS 360 ESTIMATORS.....	60
NVIDIA NeMo GUARDRAILS	60
MTEB (MASSIVE TEXT EMBEDDING BENCHMARK).....	60
MMLU (MASSIVE MULTITASK LANGUAGE UNDERSTANDING) DATASET	61
SQUAD (STANFORD QUESTION ANSWERING DATASET)	61
HOTPOTQA DATASET.....	62
MMMU (MASSIVE MULTI-DISCIPLINE MULTIMODAL UNDERSTANDING)	62
TEXTATTACK	62
NVIDIA NeMo ALIGNER.....	62
PROMPTBENCH	63
PROJECT MOONSHOT.....	64
PRODIGY.....	64
MICROSOFT PRESIDIO	64
LLM GUARD.....	65
BENCHLLM	65
ARIZE PHOENIX.....	65
GUARDRAILS AI	65
LLAMAINDEX EVALUATION TOOLS	66
GARAK	66
RAGAS.....	66
TRULENS.....	67
DEEPCHECKS	67
PROMPT FUZZER.....	67
DEEPEVAL	68
DECODINGTRUST	68
MIFLOW.....	69
TRUSTLLM BENCHMARK.....	69
CHECKLIST	70
NVIDIA NeMo	70
LLMBENCH (AGENTBENCH)	70
ARIZE PHOENIX (LLM TRACING)	71
WHYLABS	71
ADVERSARIALGLUE.....	71
APPENDICES	72
APPENDIX 1. DAGR	72
APPENDIX 2. IMPACT AND HARM ASSESSMENT.....	81
APPENDIX 3. EXAMPLES OF STATEMENTS OF CONCERN	84
APPENDIX 4. STATEMENTS OF CONCERN WORKSHEET.....	86
APPENDIX 5. RESPONSIBILITY FLOWS QUESTIONNAIRE	89
APPENDIX 6. LAWS, ETHICAL FRAMEWORKS, AND POLICIES.....	91

APPENDIX 7. PERSONAS LIST AND DESCRIPTIONS.....93
APPENDIX 8. RASCI DEFINITIONS.....95
APPENDIX 9: ACRONYM GUIDE96
APPENDIX 10: GLOSSARY97
APPENDIX 11: SUITABILITY, FEASIBILITY, ADVISABILITY ASSESSMENT 103

Introduction and Instructions

The Generative AI Version of the RAI Toolkit operationalizes the *Generative AI Guidelines & Guardrails* by offering specific questions and tools tailored to enable generative AI (GenAI) project leaders to ensure responsible and safe design, development, deployment, and use of this new technology. It shares its structure and much of its content with the RAI Toolkit MVP, released in November 2023, with some updates based on feedback the RAI team has received in the intervening year. Where appropriate, it includes new or modified content to address concerns and risks specific to GenAI.

Of particular note is the inclusion of a new tool, the Suitability, Feasibility, and Advisability Assessment. Derived from the *Report on Guidelines and Guardrails for Generative AI and Large Language Models* (April 2024), the assessment in Appendix 11 offers users a simple questionnaire to determine whether GenAI is the right technology to meet their operational needs. Using this tool as a prescreening device, AI project teams can avoid wasting time and resources pursuing a GenAI solution when other, less costly, AI or analytic techniques would prove just as—if not more—effective.

The Toolkit is intended as a technical resource to support each Component’s own governance process. Based on the risks associated with a specific project use case and the guidelines and requirements of the particular Component, users should feel free to pick and choose the most relevant pieces of this Toolkit for the user. For example, for lower-risk or experimental use cases, many of the items in the Toolkit can be skipped over. Items that are recommended for use regardless of specific use case are noted with the [Gate] tag.

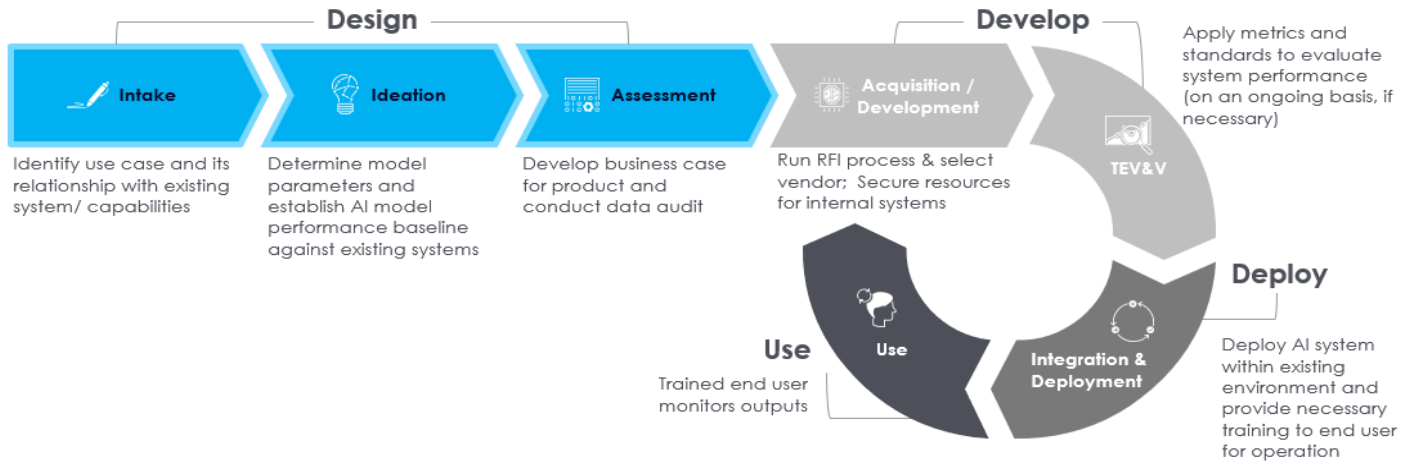
The Toolkit progresses sequentially through each stage of the product development lifecycle. The items associated with specific parts of the Guidelines & Guardrails are highlighted with footnotes as follows: ALIGNED TO: SECTION 3D - GUARDRAIL: DATA COMPLIANCE. These footnotes indicate which section of the Guidelines & Guardrails (i.e. Section 3D, in this case) is relevant to this item.

Each item of the Toolkit can also be found in the updated RAI Toolkit webapp, which has links to available tools as well as tags for each item for sorting. The GenAI Version of the RAI Toolkit is a living document and will be updated on a regular cadence.

For questions, comments, or feedback on the RAI Toolkit, please contact:
osd.pentagon.cdao.mbx.dod-rai-toolkit@mail.mil

For requests for technical or developer support on RAI issues or with technical implementation of the Toolkit, please contact the RAI Development Group (RAIDG):
cdao-raidg@groups.mail.mil

Overview of RAI Activities throughout the Product Life Cycle



Stage 1. Intake

- 1.1 Consider Previously Learned Lessons
- 1.2 Determine Relevant Laws, Ethical Frameworks, and Policies
- 1.3 Identify and Engage Stakeholders
- 1.4 Concretize the Use Case for the AI
- 1.5 Decide to Proceed to Ideation

Stage 2. Ideation

- 2.1 Define Requirements
- 2.2 Identify Risks & Opportunities / Navigate Tradeoffs
- 2.2 Weigh and Navigate Ethical Tradeoffs
- 2.3 Write Ethical Statements of Concern
- 2.4 Design to Reduce Ethical/Risk Burdens
- 2.5 Accountability, Responsibility, & Access Flows and Governance

Stage 3. Assessment

- 3.1 Assess Requirements, Statements of Concern, Mitigations, and Metrics
- 3.2 Exploratory Data Analysis
- 3.3 Update AI Suitability, Feasibility, and Advisability Assessments
- 3.4 Update Documentation

Stage 4. Development / Acquisition

- 4.1 Instrument AI to Promote Assurance
- 4.2 Update Documentation

Stage 5. TEVV

- 5.1 Test System for Robustness, Resilience, & Reliability
- 5.2 Update Documentation

Stage 6. Integration & Deployment

- 6.1 Perform Operational Testing
- 6.2 Train Users
- 6.3 Establish Incident Response Procedures
- 6.4 Auditing & Oversight Mechanisms
- 6.5 Update Documentation

Stage 7. Use

- 7.1 Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem
- 7.2 Ensure Updating and Retraining
- 7.3 Plan for System Retirement
- 7.4 Record Lessons Learned

Stage 1. Intake



1.1 Consider Previously Learned Lessons

- 1.1.1 Review similar projects compared to the use case to include incident repositories to identify RAI-related "lessons learned" applicable to the current project. [\[USE CASE REPO\]](#) [\[AI INCIDENT REPOSITORY\]](#) [\[AI INCIDENTS DATABASE\]](#)
- 1.1.2 How will novel artifacts and new lessons learned from the current project be captured and stored?

1.2 Determine Relevant Laws, Ethical Frameworks, and Policies

- 1.2.1 **[GATE]** Which legal, ethical, risk, and policy frameworks apply to this project, and how will your review and oversight process align with Component and Departmental requirements?¹ [GAMECHANGER Policy Database](#)
- 1.2.2 If your project involves the use of personally identifiable information (PII), consult with your Privacy Officer to determine if the use of the PII: 1) triggers any Privacy Act restrictions or constraints; 2) is consistent with the Fair Information Practice Principles (FIPPs) outlined in OMB Circular A-130, Appendix II; 3) requires creation or modification of a Privacy Impact Assessment (PIA) as mandated by Section 208 of the E-Government Act; 4) requires application of one or more of the Committee on National Security Systems Instruction (CNSSI) No. 1253, Privacy Overlays; 5) involves protected health information (PHI) and requires application of NIST SP 800-66, Rev. 2, Implementing the HIPAA Security Rule: A Cybersecurity Resource Guide, to the project; and, 6) necessitates use of privacy-enhancing cryptograph techniques in addition to differential privacy methods to improve the privacy posture for this project. [Appendix 2. Impact and Harm Assessment](#)
- 1.2.3 **[GATE]** How will you regularly update your ethics, policy, and legal reviews at regular intervals and/or in conjunction with changes in the application domain?
- 1.2.4 **[GATE]** Who is the responsible official(s) who will be accountable for holding the ethical, legal, and safety risks of the AI project?

1.3 Identify and Engage Stakeholders

- 1.3.1 Identify and initiate preliminary conversations with the relevant stakeholders, subject matter experts, domain experts, and operational users. [\[STAKEHOLDER MAPPING TEMPLATE\]](#)
- 1.3.2 How will the perspectives/outcomes from stakeholder engagement be surveyed/tracked throughout the product life cycle? If appropriate, highlight the process that will be used to evaluate and communicate feedback throughout the AI project.
- 1.3.3 Who will be the single authoritative leader and/or mission commander who will be responsible for engaging stakeholders or speaking externally about the project?

¹ ALIGNED TO: SECTION 3D - GUARDRAIL: DATA COMPLIANCE

1.4 Concretize the Use Case for the AI

- 1.4.1 Begin use case analysis and workflow mapping to compose a clear description of how AI supports the mission and which portions of the mission workflow are critical paths. [\[USE CASE ANALYSIS/ WORKFLOW MAPPING TOOL\]](#) [\[ROOT CAUSE ANALYSIS\]](#) [\[IMPACT GOALS TOOL\]](#) [\[FRAMING TOOL\]](#)
- 1.4.2 Identify whether multiple AI-enabled capabilities will interact with one another to support the mission. Document how they will interact with one another. [Appendix 1. DAGR](#)
- 1.4.3 **[GATE]** Have you identified the bounds of responsible/intended uses for the system? How will compromise or misuse of the system be identified?
- 1.4.4 **[GATE]** Describe your initial concept of operations (CONOPS) for the system. [\[CONOPS EXAMPLES/TEMPLATES\]](#)

1.5 Decide to Proceed to Ideation

- 1.5.1 **[GATE]** Assess the suitability, feasibility, and advisability of using Generative AI for the project. [\[GenAI Suitability, Feasibility, Advisability Assessment\]](#)
- 1.5.2 Determine the process through which stakeholders will be updated about the project and be given opportunities to provide input.

Stage 2. Ideation



2.1 Define Requirements

- 2.1.1 [GATE] Use the outputs from Stage 1 and ensure that the requirements are framed in operational terms and include a complete set of situations and conditions expected.
- 2.1.2 [GATE] Translate the operational requirements into functional requirements.
- 2.1.3 [GATE] Translate each functional requirement into technical design requirements and performance specifications.

2.2 Identify Risks & Opportunities / Navigate Tradeoffs

- 2.2.1 [GATE] Describe the project's data privacy or classification requirements, including the risks of aggregating data.
- 2.2.2 [GATE] Describe discussions that have occurred with the relevant authorizing official, and explain the process needed to receive an Authority to Operate (ATO) for this project.
- 2.2.3 Are policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of risks associated with AI in place, transparent, implemented, and validated?
- 2.2.4 [GATE] Conduct a risk assessment that includes the full scope of potential risks and document the results. The assessment should examine a scope of risks much greater than just the operation, such as risks and potential negative outcomes to society, the environment, political and economic structures, sustainability initiatives, and foreign policy partnerships and goals. Document how these tradeoffs are being navigated. [\[IMPACT ASSESSMENT TOOLS\]](#) [\[DAGR\]](#) [\[DIU WORKSHEETS\]](#) [\[NIST AI PLAYBOOK\]](#) [\[NAVIGATING TRADEOFFS TOOLKIT\]](#); see also [\[Appendix 2 for additional impact analysis questions\]](#)
- 2.2.5 How is the model's energy consumption evaluated, and what contingency plans are in place to address potential disruptions (including disruptions to the electrical infrastructure)? Is the energy consumption of the model evaluated and weighed against its expected benefits? Have smaller, more energy-efficient models been considered? [Carbon Costs Calculator](#)²
- 2.2.6 Would disruptions to the electrical infrastructure affect the model's ability to perform? What would the impact on missions and/or personnel be if power disruptions prevented the model from functioning normally?
- 2.2.7 What is the cadence for updating and revisiting risk analyses throughout the product life cycle? In what circumstances will the risk analyses need to be revisited out of cycle?

2.3 Write Statements of Concern

Statements of Concern (SOC) are RAI-related issues to be tracked across the life cycle – they may be either related to risks or potential opportunities for innovation that may be leveraged. For each SOC,

² ALIGNED TO: SECTION 6A4 - GUIDELINE: SUSTAINABILITY CONSIDERATIONS

include its estimated impact and likelihood, identify a means for establishing, updating, and tracking its priority level, and propose mitigations. SOC's can be as short as 1-2 sentence bullet points for further tracking. See Appendix 3 for SOC examples.

- 2.3.1 **[GATE] Using the legal/ethical/policy frameworks and the risks and opportunities you identified from your impact assessments in Stage 2.2, write a list of Statements of Concern. [\[STATEMENTS OF CONCERN WORKSHEET\]](#)**

2.4 Design to Reduce Ethical/Risk Burdens

- 2.4.1 Given the risk and ethical issues that surfaced from your risk assessments and SOC Worksheets, plan for how you can design your system to mitigate against these issues.
- 2.4.2 (If applicable) Begin thinking about how data and AI-enabled capabilities could potentially be leveraged to solve the SOC/ethical/risk issues that might otherwise arise from the employment or existence of the system.
- 2.4.3 How will you measure the effectiveness of these mitigations in reducing cognitive load, moral injuries, dilemmas, and other risk/ethical burdens on operational users, operational commanders, developers, and senior leaders?
- 2.4.4 How will different error types and failure modes be handled? How will error rates and failure modes be measured?
- 2.4.5 What strategies and safeguards are in place to improve the accuracy of the model and mitigate biases? Are there specific techniques to improve the accuracy of toxicity detection? How will you mitigate the risks of a GenAI model exacerbating biases and harms? [\[Human Bias Red-Teaming Toolkit\]](#), [\[DecodingTrust\]](#), [\[Tensorflow Model Remediation\]](#)
- 2.4.6 What are the potential impacts and risks of integrating the model into larger systems, and how is appropriate human oversight ensured?

2.5 Accountability, Responsibility, & Access Flows and Governance

- 2.5.1 Describe the scope of responsible use for the system by stakeholders, developers, and users of the system.
- 2.5.2 Communication behaviors: Identify what information needs to be communicated between the system and the user, and how it is currently done.³
- Does the system support adequate communication and awareness between all team members and stakeholders?
 - Is the system appropriately adaptable to users from different backgrounds, specialties, or accessibility needs?
 - Are there system characteristics, e.g. inappropriate anthropomorphism, that might encourage operators to make unwarranted assumptions about system capabilities?

³ ALIGNED TO: SECTION 5A.1 - GUIDELINE: UNDERSTANDABILITY & INTERPRETABILITY

- d) Does the operator understand the risks of inappropriate information disclosure?
- e) Does the system provide evidence or citation for information whenever possible?

2.5.3 [GATE] What degree of human involvement is needed for the system once deployed?⁴

- a) Is there a procedure for when automated decisions or activities of the system will require human approval?
- b) Are responsibilities clearly defined between the system and the human, including areas of overlap?

2.5.4 Establish accountability/responsibility flows for monitoring and addressing risks (use the Responsibility Flows Questionnaire Tool in [Appendix 5](#)). How do these responsibilities evolve at each stage of the product development life cycle? What oversight mechanisms will be established to ensure and monitor these responsibility flows and other RAI issues?⁵

⁴ ALIGNED TO: SECTION 6A1 - GUIDELINE: OVERDEPENDENCY RISKS and SECTION 6A4 - GUIDELINE: AGENTIAL HARMS

⁵ ALIGNED TO: SECTION 7C - GUIDELINE: RESPONSIBILITY FLOWS

Stage 3. Assessment



3.1 Assess Requirements, Statements of Concern, Mitigations, and Metrics

- 3.1.1 Ensure all mitigation action measures and controls have a method of being assessed and monitored throughout the life cycle. Are there any measurement gaps or limits to the precision of measurement? Will the metrics need to evolve as the system behavior changes during use (i.e. feedback loops)? How will user understanding be measured?
- 3.1.2 **[GATE] Describe the artifacts your organization requires (e.g. data ethics reviews), and your team's plans to complete them. Are all of your Statements of Concern and all aspects of your legal/ethical/policy frameworks sufficiently addressed? If not, re-conduct activities under the intake and ideation phases.**
- 3.1.3 Describe the access controls that verify the model is only accessed by those who are approved to do so, and that their access is appropriate for their specific roles. Please see [DoDD 5411](#) for guidance.
- 3.1.4 **[GATE] Revisit the performance metrics from the Feasibility Assessment (Stage 1.5.1). What additional performance metrics need to be established in light of the Ideation activities conducted in Stage 2? What evidence do you have that the proposed model will be able to meet these standards?**
- 3.1.5 Safety and error detection: identify important safety risks, indicators, and mitigations.⁶
- 3.1.6 How will you ensure the operator understands the danger and causes of AI hallucinations? [\[Garak\]](#) [\[LLamaIndex Evaluation Tools\]](#)
- 3.1.7 How will the system state (including mode of operation) be clearly communicated to the operator throughout the operation?
- 3.1.8 What means will the operator have to verify or double-check uncertain information?
- 3.1.9 What methods are employed to continuously monitor the timeliness and relevance of the dataset? How will you evaluate data representativeness to ensure sufficient coverage for the intended model tasks?
- 3.1.10 How is the model's performance monitored and adapted during use to ensure it remains suited to the operating environment? [\[Python Outlier Detection \(PyOD\)\]](#)
- 3.1.11 Is there a process in place to help identify concept drift or issues that suggest the need for model retraining?⁷ [\[Alibi Detect\]](#)
- 3.1.12 **[GATE] What are the anticipated failures? How will these be detected? Are there processes for system rollback and/or stoppage, and are these specified in the system requirements?⁸** [\[FAILURE RESOURCES\]](#) [\[ROLLBACK RESOURCES\]](#)

⁶ ALIGNED TO: SECTION 5A1 - GUIDELINE: UNDERSTANDABILITY & INTERPRETABILITY TECHNIQUES

⁷ ALIGNED TO: SECTION 3B - GUIDELINE: DRIFT

⁸ ALIGNED TO: SECTION 4A & 4B - GUARDRAIL: ADVERSARIAL ROBUSTNESS

3.2 Exploratory Data Analysis (EDA)

Where possible, EDA activities should be performed for all data that will be used in the target system. To the extent that is not possible for the original training data of foundation models, EDA is even more important for any additional data used for processes such as Finetuning or Retrieval-Augmented Generation.

- 3.2.1 What are the key properties (e.g., size, structure, data types) and attributes (e.g., accuracy, completeness, bias) of the dataset, and how do they impact its suitability for the system's intended use?
- What is the provenance of the dataset and its constituent parts? What proportion of the dataset is relevant to the system's intended use?
 - What specific selection criteria and sampling methodologies were used to ensure the dataset is representative of the intended population most relevant to the presumptive task of the GenAI model?
 - What specific steps were taken to include high-quality, authoritative sources in the training data, particularly those relevant to defense? Were sources such as military publications, defense research papers, and official government releases prioritized in the training data?
 - What measures are in place to protect personal, confidential, or classified information in the dataset?
 - How reliable are the sources of the data? Are any sources likely to become unavailable during the system's life?
 - Were any classified or sensitive government documents included, and if so, how was their inclusion managed?⁹
 - Does the dataset have any known gaps or missing data relevant to the system's intended use, and if so, how might these affect its comprehensiveness?¹⁰
- 3.2.2 **[GATE] What processes are in place to ensure that data used in model training is reliable, up-to-date, and free from errors that could compromise its validity? How current is the data? Could its currentness add elements of bias?**
- 3.2.3 How is the training data evaluated and what techniques are employed to ensure that GenAI-produced content is not used in the training runs?
- 3.2.4 How are data labels reviewed and controlled to ensure they are contextually appropriate? Is there a process or schedule for reviewing data labels and labeling decisions to ensure they fit the context they are used in or are appropriately influenced by the context in which they are used? [\[Prodigy\]](#)
- 3.2.5 How is the training data evaluated and verified to ensure its integrity, accuracy, and appropriateness for the model?¹¹ [\[IBM Fairness 360\]](#) [\[Bias and Fairness Audit Tools\]](#) [\[Word Embedding Association Tasks \(WEATs\) Method\]](#)

⁹ ALIGNED TO: SECTION 3D - GUARDRAIL: DATA COMPLIANCE

¹⁰ ALIGNED TO: SECTION 3C - GUIDELINE: DATA ATTRIBUTES

¹¹ ALIGNED TO: SECTION 5A2 - GUIDELINE: SOCIOTECHNICAL CONSIDERATIONS

- a) What steps are taken to identify and mitigate the presence of misinformation and disinformation in the training data?
 - b) What measures are in place to filter and verify the accuracy of the training data?
 - c) How is the diversity and demographic representation of the training data ensured and evaluated?¹²
 - d) Have you conducted an analysis to identify any underrepresented or overrepresented categories of information in the data?
 - e) Are there specific datasets or sources you can incorporate to address identified gaps?
- 3.2.6 Will data or feedback used to update or fine-tune the model at later stages (such as through Reinforcement Learning with Human Feedback [RLHF] or Reinforcement Learning with AI Feedback [RLAIF]) introduce any issues? How will these be mitigated? [\[NVIDIA NeMo\]](#)
- 3.2.7 **[GATE] Create Data Card with the information from this section.**¹³ [\[Hugging Face Data Card Template\]](#) [\[Aether Datasheet Template\]](#)

3.3 Update AI Suitability, Feasibility, and Advisability Assessments

- 3.3.1 Update the Suitability, Feasibility, and Advisability Assessments with new information gleaned during EDA.

3.4 Update Documentation

- 3.4.1 Update SOC's worksheet, as necessary.
- 3.4.2 What mechanisms are in place to ensure data cards remain updated as datasets evolve over time?¹⁴

¹² ALIGNED TO: SECTION 3C - GUIDELINE: DATA ATTRIBUTES

¹³ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

¹⁴ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

Stage 4. Development/Acquisition



4.1 Instrument AI to Promote Assurance

- 4.1.1 If something goes wrong with an externally procured system while it is in use, has it been established and agreed upon between AI suppliers and your organization who is responsible and who is accountable, depending on the scenario?
- 4.1.2 [GATE] Have you and the vendor budgeted for the following RAI activities, and are they incorporated into the vendor requirements?:
- a) Documentation requirements, including data/model/systems card, traceability matrix, impact assessment creation, and updating
 - b) Continuous monitoring
 - c) Model retraining and system updating
 - d) Continuous harms and impact modeling [\[NVIDIA NeMo Aligner\]](#)
 - e) Stakeholder engagement [\[Stakeholder Engagement Toolkit\]](#)
 - f) Human systems integration/human-machine teaming testing [\[RAI UX/HMT Toolkit\]](#); [\[HMT Guidebook\]](#)
 - g) User training
 - h) Assurance and Trust metrics testing [\[Trust in Autonomous Systems Test\]](#)
 - i) Routine system (and component) auditing
 - j) Sunset procedures
 - k) Uploading lessons learned into use case and incident repositories.
- 4.1.3 Ensure appropriate documentation procedures are in place: ¹⁵
- a) Will the documentation be regularly monitored and updated at each stage of product development and deployment?
 - b) Are there plans to use a traceability matrix for tracking model versions and validation and verification results?
- 4.1.4 [GATE] Ensure security procedures and requirements are defined:¹⁶
- a) For any models or services provided by third-party vendors (including data services), describe the process for conducting security assessments and audits. [\[HMT Guidebook\]](#)
 - b) What measures and safeguards do these vendors put in place to protect foundation model services against adversarial attacks? Examples include vulnerability reporting, adversarial testing or red teaming, continuous monitoring, and automated threat detection mechanisms.

¹⁵ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

¹⁶ ALIGNED TO: SECTION 4A & 4B - GUARDRAIL: ADVERSARIAL ROBUSTNESS

- c) If open-source models, datasets, and/or libraries from public repositories (e.g. HuggingFace, etc.) are adopted, describe the process for ensuring that their code base has not been compromised by adversarial tampering.
- d) What security protocols are used to ensure safe communication between the GenAI model and other components in the pipeline? Examples include role-based access control and input/output data assurance techniques. [\[NVIDIA NeMo Guardrails\]](#)
- e) What steps are taken to mitigate risks from data leaks or privacy violations in dependencies that feed data into the GenAI model?
- f) How are input sanitization and validation techniques applied to users' prompts and queries before being passed to the model? [\[NVIDIA NeMo Guardrails\]](#)
- g) How are adversarial examples incorporated during the model training or tuning process to improve GenAI model resilience? [\[TextAttack\]](#)
- h) How will you safeguard the GenAI-enabled system against the risk of data poisoning? Examples include vetting data sources, securing the end-to-end training pipeline, and using data science techniques to detect and cleanse poisoned data.
- i) What is the response plan if data poisoning is detected after the model has been trained?
- j) How is differential privacy or other privacy-preserving techniques such as secure multi-party computation [SMPC], and homomorphic encryption applied during LLM training or tuning?
- k) What processes are in place to detect and prevent adversarial attempts at prompt injections, model extraction, or replication? [\[WhyLabs\]](#)
- l) What rate-limiting policies are in place to prevent users from overwhelming the GenAI-enabled system with excessive queries or API requests?

4.1.5 Define requirements for user testing:

- a) How will operator performance be evaluated and how can it be improved?
- b) Document users' needs for model explanations given their intended workflows. Does the system support model explainability in the form of attribution of information to sources? Are the model's explanations and behaviors appropriately mapped to system behaviors and archived in a way that allows for traceability and access by key users?¹⁷ [\[Arize Phoenix LLM Tracing\]](#)
- c) Are there clear indications to users that they are interacting with an AI? What design interfaces will be used to alert the user to GenAI risks and the user's responsibilities? What warnings will alert the user if they employ poor prompting practices?¹⁸

4.1.6 Is there a process in place for identifying shortcut learning in the system? What specific techniques or tools can be used to pinpoint/uncover shortcut learning in the systems learning process?

¹⁷ ALIGNED TO: SECTION 5A1 - GUIDELINE: UNDERSTANDABILITY & INTERPRETABILITY TECHNIQUES and SECTION 5A3-4 - GUIDELINE: WATERMARKING

¹⁸ ALIGNED TO: SECTION 7A - GUIDELINE: RISKS TO HUMAN MACHINE TEAMING

- 4.1.7 How are models and model output differentiated and authenticated (e.g. watermarking, checksums, etc.)?¹⁹
- 4.1.8 Identify and understand the role of prompts in the system. Will prompt guidelines be developed and tested?²⁰ [\[PromptBench\]](#)
- 4.1.9 Are there known instances of systematic inaccuracies or misrepresentations in the model's outputs? How are systematic inaccuracies, biases, and misrepresentations in the model's outputs identified and mitigated? How does the model address source selection bias and the potential skewness introduced by dominant subsamples in the training data? [\[FairML\]](#) [\[TensorFlow Fairness Indicators\]](#) [\[Bias and Fairness Audit Tool\]](#) [\[TrustLLM Benchmark\]](#)
- 4.1.10 What measures are in place to prevent the generation of offensive or toxic language and to handle context-dependent material? Does the model differentiate between factual information and opinions or beliefs? [\[DeepEval\]](#) [\[Guardrails AI\]](#) [\[InterpretML\]](#) [\[Deepchecks\]](#)
- 4.1.11 What measures are in place to prevent the unauthorized entry, upload, or transmission of non-public DoD information, CNSI, PII/PHI, or CUI into the GenAI tool?²¹ [\[LLM Guard\]](#) [\[Microsoft Presidio\]](#)
- 4.1.12 How will appropriate DoD personnel maintain awareness of what U.S. persons data exists in the model, whether queries from a DoD Component are violating U.S. persons protections, what outputs related to U.S. person are being generated, and how those outputs are being used to support the DoD Component's activities and mission?²²
- 4.1.13 If at any point you determine that US persons-related sensitivities arise, how will you determine whether the GenAI activity should continue; and what approvals, safeguards, and oversight are appropriate to ensure the tool is used lawfully and minimizes adverse impacts on US persons?

4.2 Update Documentation

- 4.2.1 Update SOCs and data/model cards, as necessary. Consult and update DAGR to support continuous risk identification – as new risks (or opportunities) are identified.²³

¹⁹ ALIGNED TO: SECTION 5A3-4 - GUIDELINE: WATERMARKING

²⁰ ALIGNED TO: SECTION 7A - GUIDELINE: RISKS TO HUMAN MACHINE TEAMING

²¹ ALIGNED TO: SECTION 3D - GUARDRAIL: DATA COMPLIANCE and SECTION 5B - GUARDRAIL: DATA PROTECTION

²² ALIGNED TO: SECTION 6.C - GUARDRAIL: PRIVACY & CIVIL LIBERTIES PROTECTIONS

²³ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

Stage 5. TEVV



5.1 Test System for Robustness, Resilience, and Reliability

- 5.1.1 Are all parts of the stack subject to testing? Have there been unit tests of each component in isolation?
- 5.1.2 Have there been integration tests to understand how the components interact with one another within the overall system?
- 5.1.3 [GATE] How has testing established the robustness of the system and its components against adversarial attack, data/concept/model drift, data poisoning, human error and unintended or malicious use?²⁴ [\[IBM Adversarial Robustness 360 Tools\]](#)
- 5.1.4 How has testing established prompt robustness, out-of-distribution robustness, and task robustness? [\[AdversarialGLUE\]](#) [\[Checklist\]](#)
- 5.1.5 Was adversarial data integration, data augmentation, or distributionally robust optimization used? If so, document the process and outcomes.
- 5.1.6 [GATE] How has the system been tested for:
 - a) Performance? (If ground truth is not readily available, were alternate measures (extrinsic evaluation, human evaluation, diversity, and novelty metrics) used? [\[Project Moonshot\]](#) [\[MMLU dataset\]](#) [\[MMMU dataset\]](#) [\[HotpotQA dataset\]](#) [\[SQuAD\]](#) [\[MTEB\]](#)
 - b) Maintainability, including rollback procedures and proper handling of power disruptions?
 - c) Understandability of outputs?
 - d) Mechanisms to detect and prevent poisoning attacks?²⁵
 - e) The model's handling of context-dependent material, such as toxic content or offensive language. Does it perform well if toxic content is required by the use case? [\[DeepEval\]](#) [\[DecodingTrust\]](#)
 - f) Human system integration?
- 5.1.7 Were red-teaming techniques, bounties, or software tools such as fuzzing used to help identify vulnerabilities to prompt attacks? [\[Prompt Fuzzer\]](#)
- 5.1.8 How are instances of systematic inaccuracies or misrepresentations in the model's outputs identified and mitigated? [\[FairML\]](#) [\[TensorFlow Fairness Indicators\]](#) [\[Bias and Fairness Audit Tool\]](#) [\[TrustLLM Benchmark\]](#)
- 5.1.9 How is the model's ability to differentiate between factual information and opinions or beliefs tested and validated? [\[Deepchecks\]](#)
- 5.1.10 How is the model tested and validated for handling sensitive information?²⁶
- 5.1.11 If the system uses AI agents, how will you test their performance? [\[AgentBench\]](#)

²⁴ ALIGNED TO: SECTION 4A & 4B - GUARDRAIL: ADVERSARIAL ROBUSTNESS and SECTION 3C - GUIDELINE: DRIFT

²⁵ ALIGNED TO: SECTION 4A & 4B - GUARDRAIL: ADVERSARIAL ROBUSTNESS

²⁶ ALIGNED TO: SECTION 3D - GUARDRAIL: DATA COMPLIANCE and SECTION 5B - GUARDRAIL: DATA PROTECTION

5.1.12 [GATE] What are your test plan requirements for Model Finetuning? [\[Project Moonshot\]](#)

- a) How does the fine-tuned model perform (positively or negatively) on queries and tasks outside the scope of the fine-tuning objective?
- b) What are some of the processes and measures used for assessing model generalization in response to unseen data or other tasks outside the scope of the fine-tuning objective?
- c) During the fine-tuning process, how will ablation tests inform the impact of different hyperparameter settings on overall model performance?
- d) How does the model's accuracy on the validation dataset compare to its accuracy on the test dataset? [\[Mlflow\]](#)

5.1.13 [GATE] How will you measure model calibration? (See below table)²⁷

Method	Strengths	Limitations
Verbalized confidence	<ul style="list-style-type: none">• Intuitive and user friendly• Easy to implement• Applicable to all types of questions• Does not require access to underlying model	<ul style="list-style-type: none">• May require fine-tuning to return calibrated estimates
Model logits	<ul style="list-style-type: none">• Provides granular view of model output	<ul style="list-style-type: none">• Not intuitive for end-users• Most applicable to questions with fixed response options• Requires access to underlying model
Consistency-based measures	<ul style="list-style-type: none">• Easy to implement• Applicable to most types of questions• Does not require access to underlying model	<ul style="list-style-type: none">• Computationally intensive• Consistency is an indirect measure of confidence
External calibrators	<ul style="list-style-type: none">• Independent verification of model calibration	<ul style="list-style-type: none">• Technically complicated• Requires labeled training data• Requires access to underlying model

5.2 Update Documentation

5.2.1 Update SOCs, impact and risk assessments, CONOPS, security review, and data/model cards, and DAGR, as necessary.²⁸

²⁷ Technical Report Table 7

²⁸ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

Stage 6. Integration & Deployment



6.1 Perform Operational Testing

- 6.1.1 [GATE] How was operational testing performed to see whether the system works in an operational context, on the operational hardware?

6.2 Train Users²⁹

- 6.2.1 What training has been established to ensure operational end users understand system functionality, limitations, prompt techniques, selection of input data (e.g. for RAG), and other ways in which users contribute to system accuracy and effectiveness?
- 6.2.2 How are users trained to recognize and respond to system errors?
- 6.2.3 What guidelines are provided to DoD personnel to ensure they do not infringe on copyright/IP laws?
- 6.2.4 How will you evaluate whether the training was successful?

6.3 Establish Incident Response Procedures³⁰

- 6.3.1 [GATE] What is the process for flagging incidents or concerns with the system? [[Incident Response Guidance](#)]
- 6.3.2 What is the procedure for root cause analyses of system failures? [[Root Cause Analysis Toolkit](#)]

6.4 Auditing & Oversight Mechanisms

- 6.4.1 [GATE] Establish auditing procedures or oversight mechanisms for:
 - a) The overall system [[System Auditing Tools](#)]
 - b) The components in the stack (sensors, data, model, infrastructure) [[Auditing Tools](#)]
 - c) Operational users
- 6.4.2 Is it possible to construct counterfactual explanations for decisions made by the system, and has this been done as part of validation? Are chain of thought traces provided when and where possible?
- 6.4.3 Have the requirements for provenance methodologies requirements been defined? Is there a Software Bill of Materials (SBOM) that supports the particular use case (e.g., Security, compliance, and quality assurance)?

²⁹ ALIGNED TO: SECTION 4B - GUARDRAIL: REPORTING & ADVERSARIAL ROBUSTNESS, SECTION 5A1 - GUIDELINE: UNDERSTANDABILITY & INTERPRETABILITY TECHNIQUES, SECTION 7A - GUIDELINE: RISKS TO HUMAN MACHINE TEAMING, and SECTION 7B - GUIDELINE: TRAINING

³⁰ ALIGNED TO: SECTION 4B - GUARDRAIL: REPORTING & ADVERSARIAL ROBUSTNESS

6.5 Update Documentation

- 6.5.1 Update SOCs, impact and risk assessments, CONOPS, security review, data/model cards, and DAGR, as necessary.³¹

³¹ ALIGNED TO: SECTION 5A2 - GUIDELINE: DATA & MODEL CARDS

Stage 7. Use



7.1 Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem

7.1.1 [GATE] Document continuous T&E, support, and monitoring procedures to ensure performance goals continue to be met. [\[Arize Phoenix\]](#) [\[BenchLLM\]](#)

- a) Ensure the task specification remains valid.
- b) Ensure data inputs remain valid and secure.
- c) Ensure new data doesn't degrade the system.

7.1.2 Document procedures for continuous monitoring of sources of drift, changes in operational context, and human degradation/deskilling, error, and misuse. Establish procedures for mitigation of the same.

7.1.3 What incremental training processes and plans are in place?³²

7.1.4 Document procedures for continuous harms, opportunities, and impact monitoring.

- a) Ensure performance outputs and stakeholder engagement are leveraged to identify potential harms.
- b) Ensure harm and impact assessments and risk assessments, are conducted on a regular schedule.
- c) Ensure the SOCs are updated on a regular schedule.

7.1.5 [GATE] Document procedures for monitoring the system for unintended/novel uses and applications (e.g. off-label use, etc.)

7.1.6 What is your plan for updating the assessments once new functionality or features are added, new training sets out of scope from the original datasets are used, shifts or drifts have occurred, new risks have emerged, the technological landscape has changed, the broader societal or geopolitical context has changed, etc.?

7.1.7 How will you monitor the technological landscape of emerging components or systems to evaluate potential developments that may provide supplemental capabilities that could augment the performance of the system?

7.2 Ensure Updating and Retraining

7.2.1 Document mechanisms for detecting and responding to data and concept drift to ensure model performance remains optimal in dynamic environments.

7.2.2 What mechanisms, metrics, and strategies are in place to monitor, identify, and mitigate the emergence and reinforcement of unwanted norms, patterns, and biases in the model's responses?³³

³² ALIGNED TO: SECTION 7B - GUIDELINE: TRAINING

³³ ALIGNED TO: SECTION 5A2 - GUIDELINE: SOCIOTECHNICAL HARMS

- 7.2.3 [GATE] Implement processes for continuous learning, periodic retraining, or regular fine-tuning to ensure the model adapts to evolving information and maintains relevance over time. [\[TruLens\]](#) [\[RAGAS\]](#)
- 7.2.4 How is the model's performance monitored to detect early signs of collapse or degradation due to training data issues?

7.3 Plan for System Retirement

- 7.3.1 What is your sunset plan for the project end?
- 7.3.2 What are the retirement conditions for which the system will automatically be sunsetted or eclipsed?
- 7.3.3 What are the procedures for the handling of the associated hardware, software, and data to ensure they are maintained by applicable law and policy – and not repurposed in harmful ways?
- 7.3.4 How have you communicated the timelines, expectations, and criteria around the sunsetting are clear for stakeholders and operational users?

7.4 Record Lessons Learned

- 7.4.1 Report upward to AI use case repository with lessons learned so that other DoD projects can benefit from your insights. [\[RAI Use Case Repository\]](#)
- 7.4.2 Describe your plan to update the RAI Incident Repository with any near misses so that other DoD projects can benefit from your insights. PLEASE NOTE: Reporting concerns to the RAI Incident Repository does NOT automatically satisfy other reporting obligations - Reporting Locations (Must be on a DoD network to access).³⁴ [\[AI Incident Repository\]](#)
- 7.4.3 Explore other mechanisms through which to feed lessons learned or insights into policy recommendations.

³⁴ ALIGNED TO: SECTION 4B - GUARDRAIL: REPORTING & ADVERSARIAL ROBUSTNESS

The Responsible AI Tools List

Below is the RAI Tools List which contains ~100 open-source, industry-standard tools for accomplishing various RAI-related activities. These tools serve as illustrations of the kinds of tools that should be used to address the particular activities to which they were linked in the SHIELD Assessment. Many of these tools will be replaced by tools developed (or acquired) by the DoD RAI team or the RAI Working Council. For the non-DoD, open-source tools listed here, the inclusion of these open-source tools in the RAI Tools List should not be seen as an endorsement or approval of their use; DoD personnel should still go through normal approval processes before using them.

Tools that are being developed or acquired for the DoD will be indicated with a red subtitle, such as: “[**In development for the DoD**].”

Tool	Short Description or Notes	Tool Class	RAI Activities	Coding Level	Principles Mapping	Status	Documentation Link	Tool Link
RAI Use Case Repository [In Development for the DoD; LOE 1.2.2 under the RAI Strategy & Implementation Pathway]	Collection of AI use cases.	Use Case Repository	Lessons Learned; Record Lessons Learned	None	Submit the use case to the Use Case Repository to help achieve traceability (transparent and auditable documentation) and reliability (explicit uses). Compare the use case to those in the Use Case repository to help measure and demonstrate reliability (well-defined uses).			In development by CDAO RAI, MVP available FY24
AI Incident Repository [In Development; LOE 1.2.1 under the RAI Strategy & Implementation Pathway]	Collection of AI incidents and failures for review and to improve future development.	Incidents and Failure Modes	Lessons Learned; Assess Requirements, Statements of Concern, Mitigations, and Metrics; Establish Incident response Procedures; Record Lessons Learned	None	Learn from past experiences in the CDAO Incident Repository to help achieve responsibility (appropriate care), traceability (appropriate understanding of the technology and development processes), and governability (ability to avoid unintended consequences).			In development by CDAO RAI, MVP available FY24
AI Incidents Database	The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes.	Incidents and Failure Modes	Lessons Learned; Assess Requirements, Statements of Concern, Mitigations, and Metrics; Establish Incident response Procedures; Record Lessons Learned	None	Learn from past experiences in the AI Incidents Database to help achieve responsibility (appropriate care), traceability (appropriate understanding of the technology and development processes), and governability (ability to avoid unintended consequences).	Production	https://incidentdatabase.ai/about/	https://incidentdatabase.ai/
Failure Modes Resources	Document to identify threats, attacks, vulnerabilities and use the framework to plan for countermeasures. Also organizes ML failure modes and presents a framework to analyze key issues.	Incidents and Failure Modes	Lessons Learned; Assess Requirements, Statements of Concern, Mitigations, and Metrics; Establish Incident response Procedures; Record Lessons Learned	Low	Learn from past experiences in the Failure Modes Resources to help achieve responsibility (appropriate care), traceability (appropriate understanding of the technology and development processes), and governability (ability to avoid unintended consequences).	Production	https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning#how-to-use-this-document	https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
DoD AI Guide on Risk (DAGR) [In Development for the DoD; MVP Available in Appendix 1; LOE 2.1.7 under the RAI Strategy & Implementation Pathway]	Document about risks related to AI and how to calculate and mitigate risks.	RAI Checklists	Decide to Proceed to Ideation; Identify Risks & Opportunities/Navigate Tradeoffs; Responsibility Flows and Governance; AI Appropriateness Assessment	None	Keep senior leaders informed with DoD AI Guide on Risk (DAGR) to help achieve responsibility (for the development, deployment, and use of AI capabilities) and traceability (appropriate understanding of the development processes).			DAGR

IBM Design Thinking Toolkit	Activities that offer guidance to hone design thinking skills.	User Experience Design Activities	Identify and Engage Stakeholders; Determine the Use Case for the AI	None	Conduct activities using the IBM Design Thinking Toolkit to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production		https://www.ibm.com/design/thinking/page/toolkit
Design Kit Inspiration Methods	Step-by-step guide to help the design process.	User Experience Design Activities	Identify and Engage Stakeholders; Determine the Use Case for the AI	None	Conduct activities using the Design Kit Inspiration Methods to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production		https://www.designkit.org/methods.html
18F Discover Methods	Set of tools to help understand a problem and its impacts. Includes Cognitive walkthrough, Contextual inquiry, Design studio, Dot voting, Five whys, Heuristic evaluation, Hopes and fears, KJ method, Lean coffee, Stakeholder and user interviews, and Stakeholder influence mapping.	User Experience Design Activities	Identify and Engage Stakeholders; Determine the Use Case for the AI	None	Conduct activities using the 18F Discover Methods to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production		https://methods.18f.gov/discover/
Stakeholder Mapping Template	This template helps minimize confusion on who is who, clarifies responsibilities, and catalyzes a transition from strangers to collaborators. Use this template to identify stakeholders and strategize the level of involvement that each stakeholder will have.	User Experience Design Activities	Identify and Engage Stakeholders; Determine the Use Case for the AI	None	Conduct an activity using the Stakeholder Mapping Template to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production	https://www.mural.co/blog/stakeholder-mapping	https://www.mural.co/templates/stakeholder-mapping
Use Case Analysis	Walkthrough on how to develop a use case from a web design perspective.	User Experience Design Activities	Identify and Engage Stakeholders; Determine the Use Case for the AI	None	Conduct Use Case Analysis to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production		https://www.usability.gov/how-to-and-tools/methods/use-cases.html

IBM Data Privacy Toolkit	Toolkit for data type identification, privacy risk assessment, data masking and data anonymization that is exposed as a Java/Scala library and as a REST API. The toolkit consists of four main components: Type identification, Masking providers, Privacy risk assessment, Anonymization providers	Data Privacy Tools	Identify Risks & Opportunities / Navigate Tradeoffs; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, & Ecosystem	High	If input data correspond to individual people, <ol style="list-style-type: none"> 1. Apply masking techniques in the IBM Data Privacy Toolkit to help the AI capability avoid the unintended consequence of re-identifying individuals to help achieve reliability (safety of AI capabilities) and governability (ability to avoid unintended consequences). 2. Apply tests in the IBM Data Privacy Toolkit to see how well the AI capability avoids the unintended consequence of re-identifying individuals to help measure and demonstrate reliability (safety of AI capabilities) and governability (ability to avoid unintended consequences). 	Supported	https://github.com/IBM/data-privacy-toolkit/blob/main/docs/README.md	https://github.com/IBM/data-privacy-toolkit
---------------------------------	---	--------------------	--	------	---	-----------	---	---

<p>Concept of Operations Example</p>	<p>One page PDF showing an example of Concepts of Operations in the context of a helicopter.</p>	<p>Concept of Operations</p>	<p>Determine the Use Case for the AI; Revisit Documentation and Security/Roll-up into Dashboards</p>	<p>None</p>	<p>Put together a concept of operations, patterned after the Concept of Operations Example,</p> <ol style="list-style-type: none"> 1. For how the AI capability will be used, both when it is exhibiting expected behavior and when it is not, to help achieve governability (fulfill intended functions). 2. To help measure and demonstrate reliability (well-defined uses); if a detailed one can be constructed, then the use case is likely to be well-defined. 3. That includes user methods for detecting unintended AI capability behavior and deactivating or disengaging the capability to help measure and demonstrate governability (ability to disengage or deactivate deployed systems). 	<p>Production</p>	<p>https://www.dote.osd.mil/Portals/97/docs/TEMPGuide/CONOPS_Example_3.0.pdf</p>
<p>Aether Datasheet Template <i>(Additional Examples Below)</i> [Data/Model/System Card Templates are currently in Development for the DoD; MVP Release FY24; LOE 3.2.3 under the RAI Strategy & Implementation Pathway]</p>	<p>Template that includes questions that dataset creators should think through and document the answers to. Enables data provenance for multiple data sources.</p>	<p>Data Card Template</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards</p>	<p>Low</p>	<p>Fill in an Aether Datasheet Template for each data source, describing</p> <ol style="list-style-type: none"> 1. What it contains and how it was collected, to help achieve traceability (transparent and auditable documentation). 2. How it was collected, to help measure and demonstrate traceability (appropriate understanding of the development processes). 3. What it contains, to help measure and demonstrate traceability (transparent and auditable data sources). 	<p>Production</p>	<p>https://www.microsof.com/en-us/research/uploads/prod/2022/07/aether-datadoc-082522.pdf</p>

<p>Model Card Examples</p>	<p>Two model card examples for Face Detection and Object Detection.</p>	<p>Model Card Template</p>	<p>Identify Risks & Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards</p>	<p>Low</p>	<p>Compose a model card, patterned after the Model Card Examples, for each model, describing:</p> <ol style="list-style-type: none"> 1. Its technique and how it was designed, to help achieve traceability (transparent and auditable documentation). 2. Its technique, to help measure and demonstrate traceability (appropriate understanding of the technology). 3. How it was designed, to help measure and demonstrate traceability (transparent and auditable design procedures). 	<p>Production</p>	<p>https://modelcards.withgoogle.com/about</p>	<p>https://modelcards.withgoogle.com/model-reports</p>
<p>Hugging Face Data Card Template</p>	<p>Instructions on creating a dataset card. A dataset card can promote responsible usage and inform users of any potential biases within the dataset. Dataset cards help users understand a dataset's contents, the context for using the dataset, how it was created, and any other considerations a user should be aware of.</p>	<p>Data Card Template</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards</p>	<p>Low</p>	<p>Fill in a Hugging Face Data Card Template for each data source, describing:</p> <ol style="list-style-type: none"> 1. What it contains and how it was collected, to help achieve traceability (transparent and auditable documentation). 2. How it was collected, to help measure and demonstrate traceability (appropriate understanding of the development processes). 3. What it contains, to help measure and demonstrate traceability (transparent and auditable data sources). 	<p>Production</p>		<p>https://huggingface.co/docs/datasets/dataset_card</p>

Hugging Face Model Card Template	Tool, annotated template and resources to aid in model card creation. Model cards are a documentation framework for understanding, sharing, and improving machine learning models.	Model Card Template	Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards	Low	Fill in a Hugging Face Model Card Template for each model, describing: <ol style="list-style-type: none"> 1. Its technique and how it was designed, to help achieve traceability (transparent and auditable documentation). 2. Its technique, to help measure and demonstrate traceability (appropriate understanding of the technology). 3. How it was designed, to help measure and demonstrate traceability (transparent and auditable design procedures). 	Production		https://huggingface.co/blog/model-cards
Public Affairs and Communications Toolkit [In Development for the DoD]	Tools for public affairs and communications professionals.	Public Affairs and Communications Tools	Identify and Engage Stakeholders; Identify Risks & Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Plans for System Retirement	None	Use the Public Affairs and Communications Toolkit to engage and maintain transparency with stakeholders to help achieve traceability (transparent and auditable methodologies).			
Responsibility Flows Tool [In Development for the DoD; see Appendix 5 for MVP]	Tool assisting organizations define and establish roles and responsibilities for AI projects.	Responsibility Flows Tools	Responsibility Flows and Governance; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem; Plans for System Retirement	None	Delegate authority using the Responsibility Flows Tool to help achieve responsibility (for the development, deployment, and use of AI capabilities) and traceability (appropriate understanding of the operational methods). Capture, in the Responsibility Flows Tool, which commands are issued by which authorized individuals to help measure and demonstrate responsibility (for the development, deployment, and use of AI capabilities).			See Appendix 5

Planning Worksheet for DIU RAI Guidelines	Worksheet to help guide thinking and surface potential issues sooner, rather than later, to avoid unintended consequences in creating AI systems.	Impact Assessments	Identify Risks and Opportunities/Navigate Tradeoffs; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards	None	Conduct an impact assessment using the Planning Worksheet for DIU RAI Guidelines to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production		https://assets.ctfassets.net/3nanhbfr0pc/1vJvimVkijueLbJzqcaOMr/691bcab582aaf6b6de3f18b9266a6294/Planning_Worksheet_DIU-AI_Guidelines.pdf
Algorithmic Impact Assessment	Template for determining impact, including ethical considerations, impact identification and scenarios, and potential harm analysis. Originally created for the NHS.	Impact Assessments	Identify Risks and Opportunities/Navigate Tradeoffs; Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards	None	Conduct an impact assessment using the Algorithmic Impact Assessment to hone how the AI capability will be used to help achieve reliability (well-defined uses).	Production	https://www.adalovelaceinstitute.org/resource/aia-user-guide/	https://docs.google.com/document/d/12HXv7Kb4dZLnA0BkL7DiccBxoq-Slg2meBsUBq_QQQI/edit
Framework for Ethical Decision Making	Document designed as an introduction to thinking ethically and framework to aid in decision making.	Navigating Tradeoffs Tools	Identify Risks and Opportunities/Navigate Tradeoffs	None	Consult the Framework for Ethical Decision Making to iterate on the use case, optimizing it for the relevant ethical lenses, to help achieve reliability (well-defined uses).	Production	https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/introduction-to-a-framework-for-ethical-decision-making/	https://www.scu.edu/ethics/ethics-resources/a-framework-for-ethical-decision-making/

<p>IBM Fairness 360 (See below for additional tools)</p>	<p>Extensible open-source toolkit to examine, report, and mitigate discrimination and bias in machine learning models alongside explanations for metrics throughout the AI application lifecycle.</p>	<p>Group Parity Metrics; Group Parity Optimization</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If labeled training data are used and group affiliations are recorded,</p> <ol style="list-style-type: none"> 1. Use group parity metrics in IBM Fairness 360 to calculate group parity among labels to help achieve traceability (transparent and auditable data sources). 2. Use group parity optimization methods in IBM Fairness 360 to increase group parity by changing feature values in the data to help achieve equitability (deliberate steps to minimize unintended bias). <p>If group affiliations are recorded for data points, use group parity optimization methods in IBM Fairness 360 to increase group parity in the model's predicted labels by modifying the model to help achieve equitability (deliberate steps to minimize unintended bias).</p> <p>If labeled training data are used and group affiliations are recorded, use group parity metrics in IBM Fairness 360 before and after modifying the training data to calculate the change in group parity among labels to help measure and demonstrate equitability (deliberate steps to minimize unintended bias).</p> <p>If group affiliations are recorded for data points, use group parity metrics in IBM Fairness 360 before and after modifying the model to calculate the change in group parity among predicted labels to help measure and demonstrate equitability</p>	<p>Production</p>	<p>https://aif360.mybluemix.net/resources</p>	<p>https://aif360.mybluemix.net/</p>
---	---	--	--	-------------	---	-------------------	--	--

					(deliberate steps to minimize unintended bias).			
--	--	--	--	--	---	--	--	--

<p>FairML</p>	<p>Tool for auditing machine learning (ML) for bias.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in FairML to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ FairML before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply FairML to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Limited support</p>		<p>https://github.com/adebayoj/fairml</p>
----------------------	--	---------------------------------------	--	-------------	--	------------------------	--	--

<p>What-If Tool</p>	<p>Provides an easy-to-use interface for expanding understanding of a black-box classification or regression ML model. With the plugin, you can perform inference on a large set of examples and immediately visualize the results in a variety of ways. Additionally, examples can be edited manually or programmatically and re-run through the model in order to see the results of the changes. It contains tooling for investigating model performance and fairness over subsets of a dataset. The purpose of the tool is that give people a simple, intuitive, and powerful way to play with a trained ML model on a set of data through a visual interface.</p>	<p>Feature Attribution Algorithms; Counterfactual Example Generation</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>Low</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in the What-If Tool to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ the What-If Tool before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply the What-If Tool to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities). Employ the What-If Tool to generate:</p> <ol style="list-style-type: none"> 1. Counterfactual input data points that can reveal susceptibility of the AI capability to adversarial attacks to help measure and demonstrate reliability (security of AI capabilities). 2. Counterfactual input data points that can reveal logic within the AI capability that is inconsistent with policy governing the capability, hence an unintended consequence, to help measure and demonstrate governability (ability to avoid unintended consequences). 	<p>Limited support</p>	<p>https://pair-code.github.io/what-if-tool/</p>	<p>https://github.com/PAIR-code/what-if-tool</p>
----------------------------	--	--	--	------------	---	------------------------	--	--

Word Embedding Association Tasks (WEATs) Method	Paper with algorithm for enumerating biases in word embeddings.	Crowdsourcing Methods for Measuring Bias	Identify Risks and Opportunities/Navigate Tradeoffs	High	If there are potential biases that are unintended for the AI capability and are widely recognized in society, use crowdsourcing methods, such as the Word Embedding Association Tasks (WEATs) Method, to design tests to present to crowd workers before and after modifying the capability to calculate the change in bias perceived by the crowd to help measure and demonstrate equitability (deliberate steps to minimize unintended bias).	Static		https://arxiv.org/pdf/1812.08769.pdf
Bias and Fairness Audit Tool	Open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools.	Group Parity Metrics	Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security	High	If labeled training data are used and group affiliations are recorded, <ol style="list-style-type: none"> 1. Use the Bias and Fairness Audit Tool to calculate group parity among labels to help achieve traceability (transparent and auditable data sources). 2. Use the Bias and Fairness Audit Tool before and after modifying the training data to calculate the change in group parity among labels to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If group affiliations are recorded for data points, use the Bias and Fairness Audit Tool before and after modifying the model to calculate the change in group parity among predicted labels to help measure and demonstrate equitability (deliberate steps to minimize unintended bias).	Production	http://www.datasciencepublicpolicy.org/open-work/tools-guides/aequitas/	https://github.com/dssg/aequitas
Human Bias Red-Teaming Toolkit [In Development for the DoD; LOE 2.1.5 under the RAI Strategy & Implementation Pathway] (See below for additional resources)	Set of tools that helps reduce human bias in AI projects.	Human Bias Reduction Resources	Identify Risks and Opportunities/Navigate Tradeoffs; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	None	Employ the Human Bias Red-Teaming Toolkit to minimize the influence of team members' biases on the AI capability to help achieve equitability (deliberate steps to minimize unintended bias).			MVP available FY24

Cognitive Biases Resource	Infographic with 50 cognitive bases and explanations.	Human Bias Reduction Resources	Identify Risks and Opportunities/Navigate Tradeoffs; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	None	Employ the Cognitive Biases Resource to minimize the influence of team members' biases on the AI capability to help achieve equitability (deliberate steps to minimize unintended bias).	Static		https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world/
Feature Release Rollback Resource	Article on reliability of releases and deploying rollbacks.	Rollback Resources	Instrument AI to promote Assurance; Plans for System Retirement; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	Low	If the AI capability is a service that is continually available and for which new features are periodically released, use the Feature Release Rollback Resource to plan and prepare for rollbacks when new releases trigger unexpected behavior to help achieve responsibility (appropriate levels of judgment) and governability (ability to avoid unintended consequences).	Static		https://cloud.google.com/blog/products/gcp/reliable-releases-and-rollbacks-cre-life-lessons
RAI Acquisition Toolkit [In Development for the DoD; LOE 3.1.1 under the RAI Strategy & Implementation Pathway]	Set of tools for guiding and enabling potential customers on acquiring RAI products.	RAI Acquisition Tools	Instrument AI to promote Assurance; Plans for System Retirement	None				Available FY25
RAI Program Manager Review [In Development for the DoD; LOE 3.2.1 under the RAI Strategy and Implementation Pathway]	Document that establishes a guide for the review of program managers of AI projects for the use of senior leadership.	RAI Program Manager Review	Responsibility Flows and Governance	None	Equip senior leaders to decide whether to continue funding an AI capability by completing RAI Program Manager Reviews to help achieve responsibility (appropriate levels of judgment).			Available FY24
IDA Human-Machine Teaming Guidebook (See below for additional resources) [In Development for the DoD]	Paper with framework evaluating suitability and effectiveness, identifies teaming concepts, emphasizes the importance of interaction, and provides a structure for identifying and selecting appropriate measures to evaluate team effectiveness.	Human-Machine Teaming Resources	Instrument AI to promote Assurance	None	Leverage the IDA Human-Machine Teaming Guidebook to <ol style="list-style-type: none"> 1. Tailor the user interface to the intended operators to help achieve responsibility (appropriate levels of judgment) and traceability (appropriate understanding of the operational methods). 2. Tailor the user interface to the intended operators, who would disengage or deactivate if needed, to help achieve governability (ability to disengage or deactivate deployed systems). 	Static	https://www.ida.org/-/media/feature/publications/d/da/dataworks-2021-characterizing-human-machine-teaming-metrics-for-test-and-evaluation/d-21564.ashx	Available FY24

<p>MIT-LL HMT Red-Teaming Guidebook [In Development for the DoD]</p>	<p>Document establishing strategies for red-teaming HMT aspects of AI systems.</p>	<p>Human-Machine Teaming Resources</p>	<p>Instrument AI to promote Assurance</p>	<p>None</p>	<p>Leverage the MIT-LL Red-Teaming Guidebook to</p> <ol style="list-style-type: none"> 1. Tailor the user interface to the intended operators to help achieve responsibility (appropriate levels of judgment) and traceability (appropriate understanding of the operational methods). 2. Tailor the user interface to the intended operators, who would disengage or deactivate if needed, to help achieve governability (ability to disengage or deactivate deployed systems). 			<p>Available FY25</p>
<p>Human-Machine Teaming Systems Engineering Guide</p>	<p>Guide to help system developers design autonomy and automation that works in partnership with the human operator.</p>	<p>Human-Machine Teaming Resources</p>	<p>Instrument AI to promote Assurance</p>	<p>None</p>	<p>Leverage the Human-Machine Teaming Systems Engineering Guide to:</p> <ol style="list-style-type: none"> 1. Tailor the user interface to the intended operators to help achieve responsibility (appropriate levels of judgment) and traceability (appropriate understanding of the operational methods). 2. Tailor the user interface to the intended operators, who would disengage or deactivate if needed, to help achieve governability (ability to disengage or deactivate deployed systems). 	<p>Static</p>		<p>https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf</p>
<p>Trust in Autonomous Systems Test [Additional tools currently in Development for the DoD; will become available FY25; LOE 2.1.2 under the RAI Strategy & Implementation Pathway]</p>	<p>Rubric of nine criteria to evaluate a system by various metrics of trust by a user.</p>	<p>Assurance and Trust Instruments</p>	<p>Instrument AI to promote Assurance; Train Users; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>None</p>	<p>Ask users to complete the Trust in Autonomous Systems Test after they have tried out the AI capability to help measure and demonstrate traceability (appropriate understanding of the technology and operational methods).</p>	<p>Static</p>	<p>https://www.ida.org/-/media/feature/publications/p/pr/predicting-trust-in-automated-systems--validation-of-the-trust-of-automated-systems-test---toast/d-33088.ashx</p>	<p>https://osf.io/kwfmj</p>

<p>XAI Toolkit - Saliency (See Below for Additional Tools)</p>	<p>Open source, Explainable AI (XAI) framework for visual saliency algorithm interfaces and implementations, built for analytics and autonomy applications.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in the XAI Toolkit - Saliency to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ the XAI Toolkit - Saliency before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply the XAI Toolkit - Saliency to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Production</p>	<p>https://xaitk-saliency.readthedocs.io/en/latest/</p>	<p>https://github.com/XAITK/xaitk-saliency</p>
---	---	---------------------------------------	--	-------------	--	-------------------	--	--

<p>LIME</p>	<p>Local interpretable model-agnostic explanations. Used for explaining what machine learning classifiers (or models) are doing. Supports explaining individual predictions for text classifiers or classifiers that act on tables (NumPy arrays of numerical or categorical data) or images.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate LIME to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ LIME before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply LIME to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Development; Limited support</p>	<p>https://arxiv.org/abs/1602.04938</p>	<p>https://github.com/marcotcr/lime</p>
--------------------	---	---------------------------------------	--	-------------	--	-------------------------------------	--	--

<p>SHAP</p>	<p>Game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate SHAP to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ SHAP before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply SHAP to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Development</p>		<p>https://github.com/slundberg/shap</p>
--------------------	---	---------------------------------------	--	-------------	--	--------------------	--	--

<p>ExplainerDashboard</p>	<p>Used to deploy a dashboard web app that explains the workings of a (scikit-learn compatible) machine learning model. The dashboard provides interactive plots on model performance, feature importances, feature contributions to individual predictions, "what if" analysis, partial dependence plots, SHAP (interaction) values, visualization of individual decision trees, etc.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in ExplainerDashboard to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ ExplainerDashboard before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply ExplainerDashboard to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Development</p>	<p>http://explainerdashboard.readthedocs.io/</p>	<p>https://github.com/oegedijk/explainerdashboard</p>
----------------------------------	--	---------------------------------------	--	-------------	--	--------------------	--	--

<p>Diverse Counterfactual Explanations (DiCE)</p>	<p>Implements counterfactual (CF) explanations that provide information by showing feature-perturbed versions of the same scenario. Provides "what-if" explanations for model output and can be a useful complement to other explanation methods, both for end-users and model developers.</p>	<p>Counterfactual Example Generation</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>Employ Diverse Counterfactual Explanations (DiCE) to:</p> <ol style="list-style-type: none"> 1. Generate counterfactual input data points that can reveal susceptibility of the AI capability to adversarial attacks to help measure and demonstrate reliability (security of AI capabilities). 2. Generate counterfactual input data points that can reveal logic within the AI capability that is inconsistent with policy governing the capability, hence an unintended consequence, to help measure and demonstrate governability (ability to avoid unintended consequences). 	<p>Production</p>	<p>http://interpret.ml/DiCE/</p>	<p>https://github.com/interpretml/DiCE</p>
<p>ShapASH</p>	<p>Python library which aims to make machine learning interpretable and understandable by everyone. It provides several types of visualization that display explicit labels that everyone can understand. Helps data scientists understand their models easily and share their results. Allows end users to understand the decision proposed by a model using a summary of the most influential criteria. Contributes to data science auditing by displaying useful information about any model and data in a unique report.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate ShapASH to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ ShapASH before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply ShapASH to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Production</p>	<p>https://shapash.readthedocs.io/en/latest/</p>	<p>https://github.com/MAIF/shapash</p>

<p>Model Agnostic Language for Exploration and Explanation (DALEX)</p>	<p>X-rays any model and helps to explore and explain its behavior, helps to understand how complex models are working. The main function <code>explain()</code> creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers.</p>	<p>Feature Attribution Algorithms</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in Model Agnostic Language for Exploration and Explanation (DALEX) to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ Model Agnostic Language for Exploration and Explanation (DALEX) before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply Model Agnostic Language for Exploration and Explanation (DALEX) to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	<p>Production</p>	<p>https://dalex.drwhy.ai/</p>	<p>https://github.com/ModelOriented/DALEX</p>
---	---	---------------------------------------	--	-------------	---	-------------------	--	--

<p>InterpretML</p>	<p>Open-source package that incorporates state-of-the-art machine learning interpretability techniques under one roof. Helps train interpretable glassbox models and explain blackbox systems. Helps users understand model's global behavior, or understand the reasons behind individual predictions.</p>	<p>Feature Attribution Algorithms; Counterfactual Example Generation</p>	<p>Identify Risk and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in InterpretML to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences). If the features in the input data include ones along which model bias is unintended, employ InterpretML before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If the features in the input data include ones that are not intended to influence the output, apply InterpretML to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities). Employ InterpretML to:</p> <ol style="list-style-type: none"> 1. Generate counterfactual input data points that can reveal susceptibility of the AI capability to adversarial attacks to help measure and demonstrate reliability (security of AI capabilities). 2. Generate counterfactual input data points that can reveal logic within the AI capability that is inconsistent with policy governing the capability, hence an unintended consequence, to help measure and demonstrate governability (ability to avoid unintended consequences). 	<p>Development</p>	<p>https://interpret.ml/docs/getting-started</p>	<p>https://github.com/interpretml/interpret/</p>
---------------------------	---	--	---	-------------	---	--------------------	--	--

IBM Explainability 360	Extensible open-source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle.	Feature Attribution Algorithms	Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security	High	<p>If the features in the input data include ones that are not intended to influence the output, incorporate algorithms in IBM Explainability 360 to flag if weight is attributed to those features to help achieve governability (ability to detect unintended consequences).</p> <p>If the features in the input data include ones along which model bias is unintended, employ IBM Explainability 360 before and after modifying the model to calculate the change in the weight attributed to those features to help measure and demonstrate equitability (deliberate steps to minimize unintended bias).</p> <p>If the features in the input data include ones that are not intended to influence the output, apply IBM Explainability 360 to reveal if weight was attributed to those features to help measure and demonstrate reliability (effectiveness of AI capabilities).</p>	Development	https://aix360.readthedocs.io/en/latest/	http://aix360.mybluemix.net/
-------------------------------	---	--------------------------------	---	------	--	-------------	---	---

<p>Python Outlier Detection (PyOD) <i>(See below for additional tools)</i></p>	<p>Python library for detecting outlier objects in multivariate data. Includes more than 40 detection algorithms.</p>	<p>Out-of-Distribution Detection Tools</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Incorporate algorithms from Python Outlier Detection (PyOD) into the AI capability to detect out-of-distribution inputs, on which the capability's performance is not assured, to help achieve reliability (effectiveness of AI capabilities) and governability (ability to detect unintended consequences). If the AI capability preempts out-of-distribution inputs, on which its performance may exhibit unintended bias, employ Python Outlier Detection (PyOD) to evaluate how well the capability detects out-of-distribution inputs to help measure and demonstrate equitability (deliberate steps to minimize unintended bias), reliability (security and effectiveness of AI capabilities), and governability (ability to detect unintended consequences).</p>	<p>Production</p>	<p>http://pyod.readthedocs.io/</p>	<p>https://github.com/yzhao062/pyod</p>
---	---	--	--	-------------	--	-------------------	--	--

ALICE	Library for out-of-distribution detection.	Confidence Metrics; Out-of-Distribution Detection Tools	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Incorporate algorithms from ALICE into the AI capability to detect out-of-distribution inputs, on which the capability's performance is not assured, to help achieve reliability (effectiveness of AI capabilities) and governability (ability to detect unintended consequences). If the AI capability preempts out-of-distribution inputs, on which its performance may exhibit unintended bias, employ ALICE to evaluate how well the capability detects out-of-distribution inputs to help measure and demonstrate equitability (deliberate steps to minimize unintended bias), reliability (security and effectiveness of AI capabilities), and governability (ability to detect unintended consequences). Apply ALICE to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).			https://github.com/vickraj/ALICE
EQUI(NE2)	Library for uncertainty quantification.	Confidence Metrics	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Apply EQUI(NE2) to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).			https://github.com/mit-ll-responsible-ai/equine
IBM Uncertainty Quantification 360	Open-source toolkit to estimate, communicate and use uncertainty in machine learning model predictions.	Confidence Metrics	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Apply IBM Uncertainty Quantification 360 to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).	Production	https://uq360.readthedocs.io/en/latest/	https://uq360.mybluemix.net/

<p>Adversarial Patches Rearranged in Context (APRICOT) <i>(See below for additional tools)</i></p>	<p>Dataset that contains over 1000 images of printed adversarial patches in-the-wild. It is designed to be used in conjunction with the COCO dataset and COCO-trained object detection models. Created both to study the robustness of adversarial patch attacks in real-world conditions and to enable development of defensive mechanism for object detectors. Previous studies of physical adversarial objects typically tested their attacks in digital experiments or in constrained lab-like conditions. Developed to capture adversarial patches in more realistic conditions with wide variations in position, distance, lighting conditions, and viewing angles.</p>	<p>Generalization Test Datasets</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Train the AI capability on Adversarial Patches Rearranged in Context (APRICOT) to perform better on a wider array of inputs to help achieve equitability (deliberate steps to minimize unintended bias) and reliability (effectiveness of AI capabilities). Train the AI capability on Adversarial Patches Rearranged in Context (APRICOT) to better detect out-of-distribution inputs, on which the capability's performance may exhibit unintended bias, to help achieve equitability (deliberate steps to minimize unintended bias) and governability (ability to detect unintended consequences).</p>	<p>Static</p>	<p>https://arxiv.org/abs/1912.08166</p>	<p>https://apricot.mitre.org/</p>
<p>Armory Testbed</p>	<p>Testbed for running scalable evaluations of adversarial defenses. Configuration files are used to launch local or cloud instances of the Armory docker containers. Models, datasets, and evaluation scripts can be pulled from external repositories or from the baselines within this project.</p>	<p>Confidence Metrics</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Apply Armory Testbed to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).</p>	<p>Development</p>		<p>https://github.com/twosixlabs/armory</p>

Alibi Detect	Open-source Python library focused on outlier, adversarial and drift detection. The package aims to cover both online and offline detectors for tabular data, text, images and time series. Both TensorFlow and PyTorch backends are supported for drift detection.	Out-of-Distribution Detection Tools	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Incorporate algorithms from Alibi Detect into the AI capability to detect out-of-distribution inputs, on which the capability's performance is not assured, to help achieve reliability (effectiveness of AI capabilities) and governability (ability to detect unintended consequences). If the AI capability preempts out-of-distribution inputs, on which its performance may exhibit unintended bias, employ Alibi Detect to evaluate how well the capability detects out-of-distribution inputs to help measure and demonstrate equitability (deliberate steps to minimize unintended bias), reliability (security and effectiveness of AI capabilities), and governability (ability to detect unintended consequences).	Develop ment	https://docs.seldon.io/projects/alibi-detect/en/stable/	https://github.com/SeldonIO/alibi-detect
Bias Bounty Guidebook [In Development for the DoD]	Document establishing guidance on bias bounty programs.	Bias Red-Teaming Resources	Test Components for Robustness and Resilience	None	Leverage the Bias Bounty Guidebook to design and perform security tests of the AI capability to help measure and demonstrate reliability (security of AI capabilities).			Will be released in FY24
Executive Dashboard [In Development for the DoD]	Senior-level dashboards for clear understanding of RAI goals of a project.	Executive Dashboards	Decide to Proceed to Ideation; Responsibility Flows and Governance; Revisit Documentation and Security/Roll-up into Dashboards	Low	Keep senior leaders informed with the Executive Dashboard to help achieve responsibility (for the development, deployment, and use of AI capabilities) and traceability (appropriate understanding of the development processes).			MVP will be available FY25
Incident Response Guidance [In Development for the DoD; LOE 1.1.6 under the RAI Strategy & Implementation Pathway]	Interactive web application for end-user auditing.	Incident Response Guidance	Establish Incident Response Procedures	None	Heed the Incident Response Guidance when issues arise to help achieve responsibility (appropriate levels of care).			Will be released FY24

IndieLabel End-User Audit	Interactive web application for end-user auditing.	Algorithmic Auditing Tools	Revisit Documentation and Security/Roll-up into Dashboards	High	Use the IndieLabel End-User Audit to explore unexpected behavior to help measure and demonstrate traceability (auditable methodologies, data sources, and design procedures and documentation); if an audit can be performed, the AI capability is likely to be auditable.	Production; Limited support		https://github.com/StanfordHCI/indie-label
Threat Modeling Resource	Document with framework for mitigating AI/ML threats	Security Review Resources	Revisit Documentation and Security; Revisit Documentation and Security/Roll-up into Dashboards	None	Incorporate risk mitigations recommended by the Threat Modeling Resource to help achieve reliability (security of AI capabilities). Conduct a security review, guided by the Threat Modeling Resource, to help measure and demonstrate reliability (security of AI capabilities).	Static		https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml?source=recommendations
Root Cause Analysis	Template to conduct a Root Cause Analysis. A Root Cause Analysis assists in the development of a quality improvement plan for specific areas of service delivery.	Root Cause Analysis	Establish Incident Response Procedures	None	Use Root Cause Analysis to investigate unexpected behavior to help measure and demonstrate traceability (auditable methodologies, data sources, and design procedures and documentation); if a root cause analysis can be successfully conducted, the AI capability is likely to be auditable.	Static	https://www.health.state.mn.us/facilities/patientsafety/adverseevents/toolkit/index.html	https://www.in.gov/fssa/ddrs/files/008-Root-Cause-Template - BQIS_01112018.pdf

Stanford WILDS Dataset <i>(See below for additional resources)</i>	<p>Curated collection of benchmark datasets that represent distribution shifts faced in the wild. In each dataset, each data point is drawn from a domain, which represents a distribution over data that is similar in some way, e.g., molecules with the same scaffold structure, or satellite images from the same region. Includes two types of distribution shifts over domains. In domain generalization, the training and test distributions comprise disjoint sets of domains, and the goal is to generalize to domains unseen during training, e.g., molecules with a new scaffold structure. In subpopulation shift, the training and test domains overlap, but their relative proportions differ. We typically assess models by their worst performance over test domains, each of which correspond to a subpopulation of interest, e.g., different geographical regions.</p>	Generalization Test Datasets	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Train the AI capability on the Stanford WILDS Dataset to perform better on a wider array of inputs to help achieve equitability (deliberate steps to minimize unintended bias) and reliability (effectiveness of AI capabilities). Train the AI capability on the Stanford WILDS Dataset to better detect out-of-distribution inputs, on which the capability's performance may exhibit unintended bias, to help achieve equitability (deliberate steps to minimize unintended bias) and governability (ability to detect unintended consequences).	Static		https://wilds.stanford.edu/datasets/
--	--	------------------------------	---	------	---	--------	--	---

Domain Generalization Dataset List	Paper on domain generalization (DG). Contains a comprehensive literature review and formally defining DG, a review into existing methods and theories, and additional insights/discussions.	Generalization Test Datasets	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	Low	Train the AI capability on datasets in the Domain Generalization Dataset List to: <ol style="list-style-type: none"> 1. Perform better on a wider array of inputs to help achieve equitability (deliberate steps to minimize unintended bias) and reliability (effectiveness of AI capabilities). 2. Better detect out-of-distribution inputs, on which the capability's performance may exhibit unintended bias, to help achieve equitability (deliberate steps to minimize unintended bias) and governability (ability to detect unintended consequences). 	Static		https://arxiv.org/pdf/2103.02503.pdf
Checklist for ML Suitability	Checklist of 10 questions for clinicians can ask about an algorithm, but which do not require users to have knowledge of complex statistical and computational concepts.	AI Appropriateness Assessment	AI Appropriateness Assessment	None	Equip senior leaders to decide whether AI is an appropriate solution by completing the Checklist for ML Suitability to help achieve responsibility (appropriate levels of judgment).	Static		https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7871244/
GAMECHANGER Policy Database [Developed for the DoD]	Searchable database of over 50,000 DOD policy documents.	Law- and Policy-Finding Tools	Lessons Learned	None	[Missing]	Production	https://wiki.advana.data.mil/display/SDKB/GAMECHANGER+User+Guide	https://gamechanger.advana.data.mil/
Carbon Costs Calculator	Tool to estimate model carbon emissions.	Environmental Impact Estimation	Identify Risks & Opportunities/Navigate Tradeoffs	High	Apply the Carbon Costs Calculator to the AI capability to inform stakeholders of the carbon cost of using the capability to help achieve traceability (appropriate understanding of the technology).	Limited support	https://mlco2.github.io/impact	https://github.com/mlco2/impact

<p>TensorFlow Fairness Indicators</p>	<p>Tool designed to support teams in evaluating, improving, and comparing models for fairness concerns in partnership with the broader TensorFlow toolkit.</p>	<p>Group Parity Metrics</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If labeled training data are used and group affiliations are recorded, apply TensorFlow Fairness Indicators to calculate group parity among labels to help achieve traceability (transparent and auditable data sources). If labeled training data are used and group affiliations are recorded, apply TensorFlow Fairness Indicators before and after modifying the training data to calculate the change in group parity among labels to help measure and demonstrate equitability (deliberate steps to minimize unintended bias). If group affiliations are recorded for data points, apply TensorFlow Fairness Indicators before and after modifying the model to calculate the change in group parity among predicted labels to help measure and demonstrate equitability (deliberate steps to minimize unintended bias).</p>	<p>Development</p>		<p>https://github.com/tensorflow/fairness-indicators</p>
<p>TensorFlow Model Remediation</p>	<p>Python library that provides solutions for machine learning practitioners working to create and train models in a way that reduces or eliminates user harm resulting from underlying performance biases.</p>	<p>Group Parity Optimization</p>	<p>Identify Risks and Opportunities/Navigate Tradeoffs; Instrument AI to promote Assurance; Revisit Documentation and Security</p>	<p>High</p>	<p>If labeled training data are used and group affiliations are recorded, use TensorFlow Model Remediation to increase group parity by changing feature values in the data to help achieve equitability (deliberate steps to minimize unintended bias). If group affiliations are recorded for data points, use TensorFlow Model Remediation to increase group parity in the model's predicted labels by modifying the model to help achieve equitability (deliberate steps to minimize unintended bias).</p>	<p>Development</p>	<p>https://www.tensorflow.org/responsible-ai/model-remediation</p>	<p>https://github.com/tensorflow/model-remediation</p>

FoolBox	Python toolbox to create adversarial examples that fool neural networks in PyTorch, TensorFlow, and JAX.	Adversarial Attack Generation	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Employ FoolBox to: <ol style="list-style-type: none"> 1. Generate adversarial attacks and train the AI capability to be robust to them to help achieve reliability (security of AI capabilities). 2. Generate adversarial attacks and compute attack success rate to help measure and demonstrate reliability (security of AI capabilities). 	Production	https://foolbox.jonas-rauber.de/	https://github.com/bethgelab/foolbox
CounterFit	CLI that provides a generic automation layer for assessing the security of ML models.	Adversarial Attack Generation	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Employ CounterFit to: <ol style="list-style-type: none"> 1. Generate adversarial attacks and train the AI capability to be robust to them to help achieve reliability (security of AI capabilities). 2. Generate adversarial attacks and compute attack success rate to help measure and demonstrate reliability (security of AI capabilities). 	Production		https://github.com/Azure/counterfit/
SmartNoise	Differential privacy toolkit for analytics and machine learning. Uses differential privacy (DP) techniques to inject noise into data, to prevent disclosure of sensitive information and manage exposure risk.	Data Privacy Tools	Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	If input data correspond to individual people, apply masking techniques in the IBM Data Privacy Toolkit to help the AI capability avoid the unintended consequence of re-identifying individuals to help achieve reliability (safety of AI capabilities) and governability (ability to avoid unintended consequences). If input data correspond to individual people, apply tests in the IBM Data Privacy Toolkit to see how well the AI capability avoids the unintended consequence of re-identifying individuals to help measure and demonstrate reliability (safety of AI capabilities) and governability (ability to avoid unintended consequences).	Development	https://github.com/oppendp/smartnoise-sdk	https://smartnoise.org/

IBM Adversarial Robustness 360 Attacks	<p>Python library for Machine Learning Security. ART is hosted by the Linux Foundation AI and Data Foundation (LF AI and Data). ART provides tools that enable developers and researchers to defend and evaluate Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all popular machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), all data types (images, tables, audio, video, etc.) and machine learning tasks (classification, object detection, speech recognition, generation, certification, etc.).</p>	<p>Adversarial Attack Generation</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Employ IBM Adversarial Robustness 360 Attacks to:</p> <ol style="list-style-type: none"> 1. Generate adversarial attacks and train the AI capability to be robust to them to help achieve reliability (security of AI capabilities). 2. Generate adversarial attacks and compute attack success rate to help measure and demonstrate reliability (security of AI capabilities). 	<p>Production</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/ART-Attacks</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/attacks</p>
---	---	--------------------------------------	--	-------------	--	-------------------	--	--

IBM Adversarial Robustness 360 Defenses	<p>Python library for Machine Learning Security. ART is hosted by the Linux Foundation AI and Data Foundation (LF AI and Data). ART provides tools that enable developers and researchers to defend and evaluate Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all popular machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), all data types (images, tables, audio, video, etc.) and machine learning tasks (classification, object detection, speech recognition, generation, certification, etc.).</p>	<p>Out-of-Distribution Detection Tools; Input Regularization Algorithms</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Incorporate algorithms from IBM Adversarial Robustness 360 Defenses into the AI capability to:</p> <ol style="list-style-type: none"> 1. Regularize inputs to help achieve equitability (deliberate steps to minimize unintended bias). 2. Detect out-of-distribution inputs, on which the capability's performance is not assured, to help achieve reliability (effectiveness of AI capabilities) and governability (ability to detect unintended consequences). <p>If the AI capability preempts out-of-distribution inputs, on which its performance may exhibit unintended bias, employ IBM Adversarial Robustness 360 Defenses to evaluate how well the capability detects out-of-distribution inputs to help measure and demonstrate equitability (deliberate steps to minimize unintended bias), reliability (security and effectiveness of AI capabilities), and governability (ability to detect unintended consequences).</p>	<p>Production</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/ART-Defences</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/defences</p>
--	---	---	--	-------------	--	-------------------	--	--

IBM Adversarial Robustness 360 Estimators	<p>Python library for Machine Learning Security. ART is hosted by the Linux Foundation AI and Data Foundation (LF AI and Data). ART provides tools that enable developers and researchers to defend and evaluate Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all popular machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), all data types (images, tables, audio, video, etc.) and machine learning tasks (classification, object detection, speech recognition, generation, certification, etc.).</p>	<p>Confidence Metrics</p>	<p>Instrument AI to promote Assurance; Test Components for Robustness and Resilience; Operational Testing; Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem</p>	<p>High</p>	<p>Apply IBM Adversarial Robustness 360 Estimators to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).</p>	<p>Production</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/ART-Estimators</p>	<p>https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/estimators</p>
NVIDIA NeMo Guardrails	<p>Toolkit for adding programmable guardrails to LLM-based apps, ensuring safe, accurate, and on-topic conversations while preventing vulnerabilities.</p>	<p>LLM Alignment and Security</p>	<p>Instrument AI to Promote Assurance</p>	<p>Medium</p>	<p>Implement NeMo Guardrails to ensure your LLM-powered applications remain safe, accurate, and on-topic. Prevent unwanted topics and protect against vulnerabilities like prompt injections.</p>		<p>https://github.com/NVIDIA/NeMo-Guardrails</p>	<p>https://github.com/NVIDIA/NeMo-Guardrails</p>
MTEB (Massive Text Embedding Benchmark)	<p>MTEB is a benchmarking suite designed to evaluate the performance of text embedding models across a wide range of tasks and datasets.</p>	<p>Text Embedding Benchmark</p>	<p>Test System for Robustness, Resilience, and Reliability</p>	<p>Medium</p>	<p>Evaluation of text embedding models on tasks such as classification, clustering, information retrieval, and ranking, using standardized datasets.</p>		<p>https://arxiv.org/abs/2210.07316</p>	<p>https://github.com/embeddings-benchmark/mteb</p>

MMLU (Massive Multitask Language Understanding) dataset	<p>This is a large collection of mostly knowledge tests designed to provide a very broad evaluation of language models' question answering abilities. The tests are largely taken or adapted from academic (human) tests in different domains, e.g. math, medicine, philosophy.</p>	LLM Eval	Test System for Robustness, Resilience, and Reliability	Low	MMLU can be used to understand the general performance of a GenAI model and to compare performance of several models when selecting one to use.		https://paperswithcode.com/dataset/mmlu	https://huggingface.co/datasets/cais/mmlu
SQuAD (Stanford Question Answering Dataset)	<p>SQuAD is a widely used dataset for evaluating the question-answering capabilities of LLMs. It supports the evaluation of LLMs on reading comprehension tasks, where models are tasked with answering questions based on context provided in Wikipedia articles. One novel component of SQUAD is inclusion of 50,000 unanswerable questions, to test an LLMs ability to not answer questions/hallucinate where an answer is not known or not possible.</p>	LLM Eval	Test System for Robustness, Resilience, and Reliability	Low	Use SQuAD to understand a model's baseline performance at question answering, as well as its propensity to hallucinate if given an unanswerable question.		https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf	https://rajpurkar.github.io/SQuAD-explorer/

HotpotQA dataset	This is a dataset of test questions for language models focused on multi-hop questions. Multi-hop questions require more than one context passage to be returned, and in some cases require multiple sequential queries. HotPotQA questions are based on Wikipedia and each includes context passages with correct information which the RAG system can try to retrieve.	LLM Eval	Test System for Robustness, Resilience, and Reliability	Low	HotPotQA test questions can be used to understand a model's performance on multiple sequential queries.		https://nlp.stanford.edu/pubs/yang2018hotpotqa.pdf	https://hotpotqa.github.io/
MMMU (Massive Multi-discipline Multimodal Understanding)	This dataset is similar to the Massive Multitask Language Understanding dataset, but extended to include questions that have a multimodal component, mostly images.	LLM Eval	Test System for Robustness, Resilience, and Reliability	Low	MMMU can be used to understand the general performance of a GenAI model and to compare performance of several models when selecting one to use.		https://arxiv.org/abs/2311.16502	https://mmmu-benchmark.github.io/
TextAttack	TextAttack is an open-source Python framework for adversarial attacks, data augmentation, and training of NLP models.	Adversarial attack Generation	Instrument AI to Promote Assurance	High	Use TextAttack to evaluate the robustness of your NLP models by generating adversarial examples and enhancing model resilience through adversarial training and data augmentation.		https://github.com/QData/TextAttack	https://github.com/QData/TextAttack
NVIDIA NeMo Aligner	Scalable toolkit supporting state-of-the-art alignment techniques like RLHF, DPO, and SteerLM for safe, helpful, and reliable LLMs.	LLM Alignment and Safety	Instrument AI to Promote Assurance	High	Apply NeMo Aligner to align LLMs with human values using techniques like RLHF and DPO. Fine-tune models for safety, reliability, and performance at large scale.		https://github.com/NVIDIA/NeMo-Aligner	https://github.com/NVIDIA/NeMo-Aligner

<p>PromptBench</p>	<p>PromptBench is a benchmarking platform designed to evaluate the effectiveness of prompts across different large language models (LLMs). It enables researchers and developers to compare how well various prompt designs perform on tasks like text generation, summarization, and question answering. By providing standardized evaluations, PromptBench helps improve prompt engineering strategies and optimize model outputs for specific tasks.</p>	<p>LLM Eval; Red Teaming</p>	<p>Instrument AI to Promote Assurance</p>	<p>Low</p>	<p>Use PromptBench to track and compare prompting designs to provide the best possible inputs to your LLM.</p>		<p>https://promptbench.readthedocs.io/en/latest/examples/basic.html</p>	<p>https://github.com/microsoft/promptbench</p>
---------------------------	---	------------------------------	---	------------	--	--	--	--

Project Moonshot	Project Moonshot is one of the world's first Large Language Model (LLM) Evaluation Toolkits, designed to integrate benchmarking, red teaming, and testing baselines. It helps developers, compliance teams, and AI system owners manage LLM deployment risks by providing a seamless way to evaluate their applications' performance, both pre- and post-deployment. This open-source tool is hosted on GitHub and is currently in beta.	LLM Eval; Red Teaming	Test System for Robustness, Resilience, and Reliability		Use the toolkit from Project Moonshot to benchmark and compare performance of your LLM across multiple use cases and through the finetuning process.		https://github.com/aiverify-foundation/moonshot	https://aiverifyfoundation.sg/project-moonshot/
Prodigy	Prodigy is an annotation tool designed to help developers and data scientists label large datasets efficiently.	Annotation Tool	Exploratory Data Analysis	Medium	Use Prodigy to rapidly annotate and curate datasets for training LLMs, leveraging active learning to maximize labeling efficiency and improve model accuracy.		https://prodi.gy/docs/large-language-models	https://prodi.gy/docs/large-language-models
Microsoft Presidio	Presidio is an open-source data protection and privacy tool developed by Microsoft. It detects and anonymizes sensitive personal data (PII) in text and speech.	Privacy Tool	Instrument AI to Promote Assurance	Medium	Use Presidio to detect and anonymize sensitive personal data in both structured and unstructured datasets, ensuring data privacy and regulatory compliance.		https://microsoft.github.io/presidio/	https://microsoft.github.io/presidio/

LLM Guard	Open-source toolkit for securing LLM interactions, detecting prompt injections, and handling PII. Provides input and output scanners for protection.	LLM Alignment and Security	Instrument AI to Promote Assurance	Medium	Incorporate LLM Guard to secure LLM applications against prompt injections and ensure data privacy. Deploy input/output scanners for PII and adversarial attack prevention.		https://github.com/protectai/llm-guard	https://github.com/protectai/llm-guard
BenchLLM	Open-source tool for testing, evaluating, and benchmarking LLM-powered applications, with custom evaluation strategies.	LLM Eval	Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Incorporate BenchLLM to automate the evaluation of LLM models, ensuring semantic consistency across test cases. Use it to track model regressions and continuously improve LLM performance by comparing outputs over time.		https://github.com/v7labs/benchllm	https://benchllm.com
Arize Phoenix	Open-source platform for LLM observability, tracing, and evaluation to optimize model performance and track issues during development and production.	LLM Alignment and Evaluation	Perform Continuous Monitoring of the System and its Use, Context, and Ecosystem	High	Leverage Phoenix to trace, evaluate, and optimize LLM applications throughout their lifecycle. Apply it to detect performance bottlenecks and manage prompt iterations, ensuring that your LLM continues to meet performance expectations during both development and production.		https://docs.arize.com	https://phoenix.arize.com
Guardrails AI	Open-source framework for safety, security, and performance checks in LLM applications, preventing hallucinations and ensuring factual outputs.	LLM Alignment and Security	Instrument AI to Promote Assurance	High	Leverage Guardrails AI to define and enforce rules that prevent LLMs from generating hallucinations or inappropriate content. Use the framework to implement safety checks and compliance standards within any LLM-powered application.		https://www.guardrailsai.com/docs/concepts/hub	https://www.guardrailsai.com

LlamaIndex Evaluation Tools	LlamaIndex offers a set of evaluation metrics along with other features of its RAG construction toolkit. The toolkit will automatically generate questions from context to supplement human-written tools, and does automated RAG response and response scoring via LLM (can be two different LLMs).	LLM Alignment and Evaluation	Assess Requirements, Statements of Concern, Mitigations, and Metrics	Medium	Use evaluation tools to understand whether your RAG application will generate the right responses for your workflow		https://docs.llamaindex.ai/en/stable/optimizing/evaluation/evaluation/	https://docs.llamaindex.ai/en/stable/
Garak	Garak is a platform designed to detect and mitigate LLM hallucinations, helping developers create safer and more accurate language models. It provides tools for analyzing and identifying hallucinations and vulnerabilities in model outputs.	LLM Alignment and Evaluation	Assess Requirements, Statements of Concern, Mitigations, and Metrics	Medium	Use Garak to monitor and prevent hallucinations in LLM-generated text, ensuring that the outputs are accurate, safe, and aligned with user expectations.		https://docs.garak.ai/garak/overview/our-features	https://docs.garak.ai/garak/overview/our-features
RAGAS	Framework for evaluating Retrieval Augmented Generation (RAG) pipelines, focusing on answer relevance, faithfulness, and recall.	LLM Alignment and Evaluation	Ensure Updating and Retraining	High	Apply RAGAS to systematically evaluate the relevance and faithfulness of answers produced by Retrieval Augmented Generation (RAG) pipelines. Integrate RAGAS in your CI/CD pipeline to maintain high accuracy in response generation, especially for retrieval-heavy applications.		https://docs.ragas.io/en/latest/	https://ragas.io

TruLens	Evaluation and tracking framework for LLM apps using feedback functions to measure quality metrics like relevance and safety.	LLM Alignment and Evaluation	Ensure Updating and Retraining	High	Integrate TruLens feedback functions to monitor and improve key metrics such as groundedness, safety, and answer relevance in LLM applications. Use it to compare different model versions and iteratively enhance LLM performance based on contextual evaluations.		https://www.trulens.org	https://www.trulens.org
Deepchecks	End-to-end LLM evaluation and monitoring platform for testing model performance across multiple stages, from pre-deployment to production.	LLM Alignment and Evaluation	Instrument AI to Promote Assurance; Test System for Robustness, Resilience, and Reliability	High	Utilize Deepchecks to automate the validation and monitoring of LLM models across different deployment stages. Apply it to run pre-deployment checks, ensuring that your model outputs remain factual and aligned with business goals post-deployment.		https://llmdocs.deepchecks.com/docs/what-is-deepchecks	https://deepchecks.com/llm-evaluation
Prompt Fuzzer	Interactive tool to assess the security of your GenAI application's system prompt against various dynamic LLM-based attacks.	LLM Alignment and Security	Test System for Robustness, Resilience, and Reliability	High	Use Prompt Fuzzer to improve your system's reliability in the face of attacks such as jailbreak, prompt injection, or system prompt extraction.	Production	https://github.com/prompt-security/ps-fuzz/blob/main/README.md	https://github.com/prompt-security/ps-fuzz

<p>DeepEval</p>	<p>DeepEval is one of many packages that helps conduct assessments of LLM systems, particularly RAG-based systems. https://docs.confident-ai.com/docs/guides-rag-evaluation Metrics included: Summarization, Answer Relevancy, Faithfulness (of answer to returned context), Contextual Precision and Contextual Recall (compare RAG-retrieved content to given ground truth), Tool Correctness, Hallucination, Bias, Toxicity. Includes some metrics including Conversational Completeness, Conversational Relevancy, Knowledge Retention.</p> <p>Similar packages include: TruLens, RAGAS, InspectorRAGet, Microsoft's RAG Experiment Accelerator, Tonic, Decoding Trust, as well as evaluation functions that are included with LlamaIndex, LangChain, MLFlow.</p>	<p>LLM Alignment and Evaluation</p>	<p>Instrument AI to Promote Assurance; Test System for Robustness, Resilience, and Reliability</p>	<p>Medium</p>	<p>Use DeepEval to implement continuous evaluation of LLM applications through custom and out-of-the-box metrics. Integrate DeepEval into your CI/CD pipeline to monitor LLM performance over time and improve key outputs such as accuracy, relevance, and toxicity prevention.</p>		<p>https://docs.confident-ai.com</p>	<p>https://github.com/confident-ai/deepeval</p>
<p>DecodingTrust</p>	<p>DecodingTrust is a benchmarking platform aimed at evaluating the trustworthiness of LLMs.</p>	<p>LLM Alignment and Evaluation</p>	<p>Design to Reduce Ethical Burdens/Risk; Test System for Robustness, Resilience, and Reliability</p>	<p>Medium to High (Python-based, with APIs for integrating evaluations into existing LLM workflows)</p>	<p>https://decodingtrust.github.io/</p>		<p>https://arxiv.org/abs/2306.11698</p>	<p>https://decodingtrust.github.io/</p>

Mlflow	<p>Comprehensive platform for tracking, managing, and deploying models in the machine learning lifecycle, supporting multiple ML and LLM libraries. MLFlow offers automated RAG evaluation functions as part of its larger automation framework. It is similar to other tools in that it will automatically generate questions from context to supplement human-written tools, and does automated RAG response and response scoring via LLM (can be two different LLMs). MLFlow also makes it easy to put together and gather data from multiple pre-planned evaluation runs.</p>	LLM Alignment and Evaluation	Test System for Robustness and Resilience	High	<p>Use MLflow to track experiments, manage LLM models, and streamline their deployment. Employ it to maintain consistency and reproducibility in LLM development while ensuring traceability through every stage of the model lifecycle.</p>		https://mlflow.org/docs/latest/index.html	https://mlflow.org
TrustLLM Benchmark	<p>Comprehensive framework for assessing the trustworthiness of LLMs across six dimensions, including truthfulness, safety, fairness, and robustness.</p>	LLM Alignment and Evaluation	Instrument AI to promote Assurance; Test System for Robustness and Resilience	Medium	<p>Use TrustLLM to evaluate LLMs for safety, fairness, and privacy. Ensure models meet industry standards for ethical AI by benchmarking performance across datasets.</p>		https://trustllmbenchmark.github.io/TrustLLM-Website	https://trustllmbenchmark.github.io/TrustLLM-Website/index.html

Checklist	Checklist is a tool for testing NLP models using a list of task-specific capabilities. Checklist supports automated and semi-automated synthetic data generation, robustness testing with data perturbations, minimum functionality testing, invariance testing, and directed expectation testing.	LLM Alignment and Evaluation	Test System for Robustness, Resilience, and Reliability	Medium	Use Checklist to systematically evaluate and debug your NLP models by testing specific capabilities such as robustness to paraphrasing and handling of negations.		https://github.com/marcotcr/checklist	https://github.com/marcotcr/checklist
NVIDIA NeMo	Cloud-native framework for developing, customizing, and deploying LLMs, multimodal models, and speech AI applications.	LLM Alignment and Security	Exploratory Data Analysis	High	Use NeMo for developing and fine-tuning LLMs across distributed systems. Implement alignment techniques like RLHF to ensure model safety and scalability.		https://github.com/NVIDIA/NeMo	https://github.com/NVIDIA/NeMo
LLMBench (AgentBench)	LLMBench (AgentBench) is a benchmarking platform specifically designed to evaluate the performance of LLM-based agents in complex, real-world tasks. It provides a structured environment where LLM agents can be assessed on decision-making, reasoning, and multi-step task execution across a variety of scenarios. LLMBench helps developers identify limitations and optimize LLMs for autonomous task management,	LLM Alignment and Security	Test System for Robustness, Resilience, and Reliability	Medium	Run LLMBench to evaluate the ability of LLMs to act as autonomous agents in complex environments. Use it to benchmark different models, identify limitations in long-term reasoning, and optimize LLMs for real-world agent-based tasks.		https://llmbench.ai/agent	https://llmbench.ai/agent

Arize Phoenix (LLM Tracing)	Arize Phoenix provides detailed tracing for LLMs, allowing developers to visualize and troubleshoot the outputs of language models in production.	LLM Alignment and Evaluation	Instrument AI to Promote Assurance	Medium	Use Arize Phoenix to trace the flow of prompts and responses in LLM systems, ensuring transparency and troubleshooting performance issues.		https://docs.arize.com/phoenix/tracing/llm-traces	https://docs.arize.com/phoenix/tracing/llm-traces
WhyLabs	AI observability and security solution focusing on real-time monitoring of LLM deployments, detecting risks like hallucinations or prompt injections.	LLM Alignment and Security	Instrument AI to Promote Assurance	Medium	Deploy WhyLabs for real-time security monitoring of LLM deployments, using it to detect prompt injections, hallucinations, and other security vulnerabilities. Use WhyLabs to safeguard data privacy and ensure regulatory compliance in sensitive environments.		https://docs.whylabs.ai	https://whylabs.ai
AdversarialGLUE	AdversarialGLUE is an extension of the GLUE benchmark, designed to evaluate the robustness of natural language understanding models against adversarial examples.	Adversarial attack Generation	Test System for Robustness, Resilience, and Reliability	Medium	Use AdversarialGLUE to benchmark your model's robustness and identify potential vulnerabilities to adversarial perturbations in natural language processing (NLP) tasks.		https://adversarialglue.github.io/	https://adversarialglue.github.io/

Appendices

Appendix 1. DAGR

Defense AI Guide on Risk (DAGR)

The Responsible Artificial Intelligence (RAI) Defense AI Guide on Risk (DAGR) is intended to provide DoD AI stakeholders with guiding principles to promote improved trustworthiness, effectiveness, responsibility, risk mitigation, and operations that align to the DoD AI Ethical Principles, NIST AI Risk Management Framework (AI RMF), best practices, and other governing Federal and DoD guidance.

There are seven guiding dimensions of risk within the NIST AI RMF. According to the AI RMF, units should strive to ensure AI platforms and applications are (1) valid and reliable, (2) safe, (3) secure and resilient, (4) explainable and interpretable, (5) privacy-enhanced, (6) fair – with harmful bias managed, and (7) accountable and transparent. Further expansion of the seven guiding dimensions of AI risk can be found in the latest NIST AI RMF document. It is important to note that the series of questions provided by DAGR are not intended to serve as go/no-go criteria or authority-to-operate (ATO), but rather, to serve as a guide to promote risk thought and assessment with AI capabilities.

Several approaches to examine risk and impacts are used within private and public domains, such as the Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis or Political, Economic, Social, and Technological (PEST) analysis. In support of additional societal concerns, and federal regulations and guidance (such as Presidential Executive Orders and The Department of Defense elevating climate change as a national security priority), DAGR expands upon the PEST model to provide a novel approach referred to as a STOPES analysis.

A STOPES analysis examines the Social, Technological, Operational, Political, Economic, and Sustainability (STOPES) factors.

Stopes refers to a mining concept of excavating a series of steps or layers, and the STOPES acronym is fitting to the concept of risk management within AI due to the significant implications that may arise from an AI

capability and the requirement to analyze and explore layers of impacts and risk across multiple disciplines. As AI stakeholders refer to the guidance provided here, it is important to incorporate a STOPES analysis to fully realize the risks of AI capabilities and mitigate accordingly.

Risk is a dynamic concept, which is realized in context, therefore, it is prudent to evaluate risk not only throughout the operational window of an AI capability but across the entire lifecycle. In addition to being dynamic, risk may be realized due to relationships with other AI capabilities or the environment it operates within. It is the responsibility of the decision maker, commander, and warfighter to appropriately assess risk throughout the lifecycle.

The DAGR deliberately utilizes the mnemonic to symbolize how – contrary to a common objection – RAI does not impede warfighter effectiveness but instead promotes increased capability to the warfighter and support for commander intent, all within acceptable risk parameters. Additionally, when performed appropriately, RAI earns the confidence of the American public, industry, academia, allies, partners, and the broader AI community to sustain our technological edge and capability from the boardroom to the battlefield.

The content of this document reflects recommended practices. This document is not intended to serve as or supersede existing regulations, laws, or other mandatory guidance.

1. DAGR Intended Audience and Profile

According to the NIST AI RMF, use-case profiles are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on requirements, risk tolerance, and resources of the framework user. For example, an AI RMF *operational commander/decision-maker* profile or AI RMF *cybersecurity* profile may have different risks to highlight and address throughout the AI capability lifecycle.

The DAGR's primary purpose is to highlight applicable RAI concepts, holistically guide risk evaluation, and

provide an abstracted risk model to mitigate risk of AI capabilities while promoting responsibility and trust. Future expansions of the DAGR will include risk highlights, questions, and mitigations for appropriate profiles. For example, an AI RMF *operational commander/decision-maker* profile guidance will be produced that delineates appropriate risk factors that must be evaluated at this level of responsibility while also abstracting more technical and detailed factors that will be addressed by other profiles but must still be considered by the *operational commander/decision maker*. Another example may be an AI RMF *data science* profile that would conduct detailed risk evaluations with an emphasis on data, such as computational biases, but the *operational commander/decision-maker* profile must ensure it is addressed and mitigated appropriately without necessarily focusing on the technical details.

2. Introduction and Key Concepts

The DoD has made significant progress in establishing policy and strategic guidance for the adoption of AI technology since the 2018 National Defense Strategy (NDS) that highlighted the Secretary of Defense's recognition of the importance of new and emerging technologies, including AI. Since then, the DoD has released the "DoD AI Ethical Principles", which highlights that the following principles must apply to all DoD AI capabilities, encompassing both combat and non-combat applications:

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

Equitable: *Deliberate steps must be taken to minimize unintended bias in AI capabilities.*

Traceable: *DoD AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development process, and operational methods applicable to AI capabilities, including transparent and auditable methodologies, data sources, and design procedure and documentation.*

Reliable: *DoD AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire lifecycle.*

Governable: *AI capabilities will be designed and engineered to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage/deactivate deployed systems that demonstrate unintended behavior.*

In addition to the established DoD AI Ethical Principles, the National Institute of Standards and Technology (NIST) released the AI RMF in early 2023. Within this framework, stakeholders should ensure that AI platforms and applications mitigate risk to an acceptable level across the seven guidelines. For the remainder of this document, AI systems, models, platforms, and applications will be referred to as AI capabilities from here on out. The seven guidelines are:

Valid and Reliable: *Factors to confirm, through objective evidence, that the requirements for a specific intended use or application have been fulfilled. Also included are factors to evaluate the ability of a capability to perform as required, without failure, for a given time interval, under given conditions.*

Safe: *Factors that ensure that under defined conditions, AI capabilities should not lead to a state in which human life, health, property, or the environment are endangered.*

Secure and Resilient: *Factors to evaluate AI capabilities and an ecosystem's ability to withstand unexpected adverse events or unexpected changes in their environment or use. This includes factors related to robustness, maintainability, and recoverability.*

Accountable and Transparent: *Factors related to the extent to which information about an AI capability and its outputs are available to stakeholders interacting with an AI capability.*

Explainable and Interpretable: *Factors related to the extent to which the operational mechanisms of an AI capability can be explained, and the meaning of an AI capability's output is as designed for operational purposes (interpretability).*

Privacy-Enhanced: *Factors related to the norms and practices that contribute to the safeguarding of human autonomy, identity, and dignity.*

Fair – with Harmful Bias Managed: Factors related to the concerns for equality and equity due to harmful bias and discrimination. Computational factors are generally focused on, but emphasis must be placed on human and systemic biases. **Computational bias** may occur when a sample is not accurate or representative of the population/subject in question. **Human bias** may occur due to systemic errors in human thought and perception. **Systemic bias** may occur from beliefs, processes, procedures, and practices that may result in advantages and disadvantages for various social groups.

It is vital to the concept of RAI that both the DoD AI Ethical Principles and AI RMF be considered together to capture a greater risk picture.

3. AI Risk Relationship Dynamics

When exploring and assessing AI risk, there are several risk relationship dynamics to consider before the deployment of an AI capability.

AI risk has a **shifting** dynamic, meaning that throughout the lifecycle of an AI capability, overall risk may shift and become more or less impactful.

AI risk *may* have a **bidirectional** dynamic with the environment. This implies that an AI capability may influence the environment it operates within, and the environment may influence the AI capability. Figure 1 represents this potential bidirectional dynamic between the AI capability and the environment.

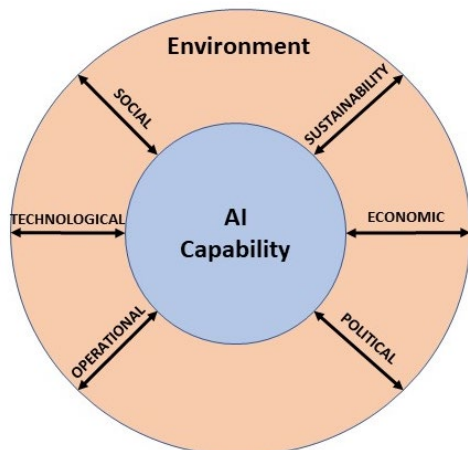


Figure 1. Bi-directional relationship of AI capability and environment.

Risks between AI capabilities may be **interconnected** because of a dependency between AI capabilities. This implies that residual risk for an AI capability may

affect the risk of another AI capability if they share a dependency. The risk dynamic between dependent AI capabilities does not necessarily have to be bi-directional, meaning that if a dependency between two capabilities exists, it does not have to be a bi-directional dependency. This risk dynamic **may only** be realized if there is a dependency between two capabilities. A dependency is when one capability requires another capability to perform its designated function.

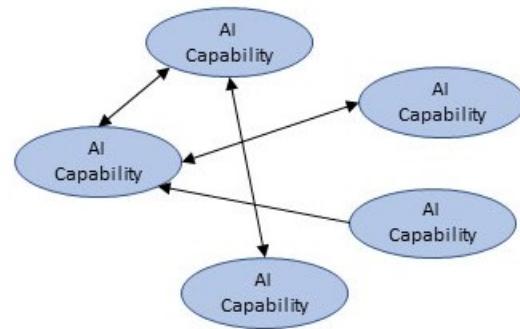


Figure 2. Interconnected relationship of AI risk between capabilities.

4. Risk Factors, Risk Assessment, and a Hierarchy of AI Risk

During the lifecycle of an AI capability, AI stakeholders must evaluate the **shifting, interconnected, and bidirectional** dynamics of risk across the social, technological, operational, political, economic, and sustainability (STOPES) factors. Within this section, the concepts of STOPES and high-level risk perspectives will be presented to guide further in-depth risk discussion and assessment.

Important note: Before reviewing the STOPES factors, it is necessary to highlight that supply chain risks may be evaluated across several of the STOPES factors. Common supply chain risks include poor contractor and/or supplier performance, labor shortages, funding, geopolitical, cyber, environmental, and reputational risks, to highlight a few.

Reviewing the above list of supply chain risks as an example, it can be surmised that cyber supply chain risks related to software and hardware would reside within technological factors, but poor contractor and/or supplier performance, labor shortages, and funding would be categorized within economic factors of the STOPES analysis.

Social Factors: Factors related to community, social

support, income, education, race and ethnicity, employment, and social perceptions.

Technological Factors: Factors related to the organizational impacts of a technological capability being inoperable, compromised, or operating incorrectly, and appropriate supply chain risks related to technology and security.

Operational Factors: Factors that may result in adverse change in resources resulting from operational events such as military (combat and non-combat) operations, operations inoperability or incorrectness of internal processes, systems, or controls, to also include external events and appropriate supply chain risks related to operations. Operational factors also include reputation, legal, ethical, and human-machine interaction, and the corresponding feedback loop of this interaction.

Political Factors: Factors related to government policy, changes in legislation, political climate, and international relations.

Economic Factors: Economic factors that may influence the organization, such as access to funding, acquisition processes and vehicles, labor costs and workforce skill, market conditions, and appropriate supply chain risks related to economics.

Sustainability Factors: Factors related to human-, environmental-, social-, and economic sustainability. The intersection and balance of environment, economy, and social equity support sustainability initiatives. It is important to note that the social sustainability and economic sustainability factors are specialized topics within the aforementioned economic and social categories and are recommended to be highlighted here. Factors include concepts related to climate change, the environment, energy usage, social responsibility, human security, and appropriate supply chain risks related to sustainability.

When evaluating risk across the STOPES factors, we recommend utilizing the DISARM hierarchy of AI risk. DISARM is an acronym for data, infrastructure, security, accountability, resources, and model operation. The DISARM hierarchy is intended to serve as an abstracted model to prioritize the evaluation and mitigation of risk of an AI capability. It is important to note that addressing the risk at a lower level of the hierarchy does not mean that a

stakeholder can move up the hierarchy and ignore risk factors of lower levels at later intervals of the AI capabilities lifecycle.

Within the DISARM hierarchy of AI risk, AI stakeholders should strive to mitigate risks associated with data first since the collection, cleaning, and labeling of data are critical in reducing bias and promoting validity, reliability, effectiveness of AI capabilities, and mitigating risk at the higher levels of the hierarchy.

Continuing along the hierarchy, the AI stakeholder will continue evaluating risks and factors associated with infrastructure, security, accountability, resources, and model operations, in order. Without adequately mitigating risk at the lower levels, AI stakeholders cannot expect risk to be appropriately mitigated at the higher levels, and for the AI capability to operate as intended. Figure 3 depicts the DISARM hierarchy of AI risk.

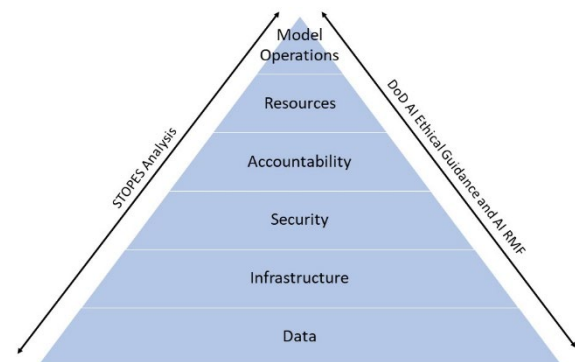


Figure 3. DISARM Hierarchy of AI Risk.

Below is a description of each layer of the DISARM hierarchy of AI risk, but this is not intended to be an all-inclusive list.

Data: The foundation of the DISARM hierarchy, is the necessity to evaluate factors associated with data, which include ensuring that data is collected appropriately (manual and/or automated collection), data is accurate, integrity is maintained, and biases are mitigated.

Infrastructure: Factors related to the infrastructure in which data and capabilities traverse or interact, including any sensors. Infrastructure factors also include elements that may impact the availability of data to support AI capabilities.

Security: Factors related to the security, resiliency,

and privacy of data, technology, and people. Security factors include elements that may impact the confidentiality, integrity, availability, authentication, and non-repudiation of AI capabilities. Several cybersecurity risk management frameworks are in existence, and selection is dependent on applicable authorities. For example, utilizing the NIST Cybersecurity Risk Management Framework may be appropriate.

Accountability: Accountability is a prerequisite to transparency and includes accountability factors to AI capabilities, society, DoD units/forces, and partners and allies. Accountability also involves evaluating factors related to an AI capability being valid, reliable, explainable, and interpretable. By addressing these risk factors, AI stakeholders are able to promote an acceptable level of trust and confidence in operations using an AI capability.

Resources: Factors related to people, equipment, or ideas available to respond to a threat or hazard to an AI capability.

Model Operations: Factors related to the correct operation of a model or AI capability during the operational window.

5. AI Capability and Risk Lifecycle

According to the DoD Responsible AI Strategy and Implementation Pathway, the AI product lifecycle consists of the iterative activities of design, develop, deploy, and use. Equally iterative is the need to evaluate and address risk throughout the AI lifecycle and across activities.

For simplicity, Figure 4 combines Deploy and Use within the same phase. The AI capability lifecycle consists of the following phases:

Design: Consists of the activities of intake, ideation, and assessment. Within the design, it is necessary to identify the use case, and its relationship with other existing capabilities, collect requirements, finalize desired outcomes and objectives, conceptualize and design the AI capability, and evaluate available resources for the problem.

Develop: According to the Defense Innovation Unit, this phase refers to the iterative process of writing and evaluating the AI capabilities program code. Also, all development stakeholders should focus on the

management of data models, continuous monitoring, output verification, audit mechanisms, and governance roles.

Deploy/Use: This phase encompasses the processes of utilizing the AI capability during the operational window, providing training to end-users for operations, and monitoring outputs.

Important note: Various STOPES factors and risks may be less or more relevant during different phases of the AI capability lifecycle.

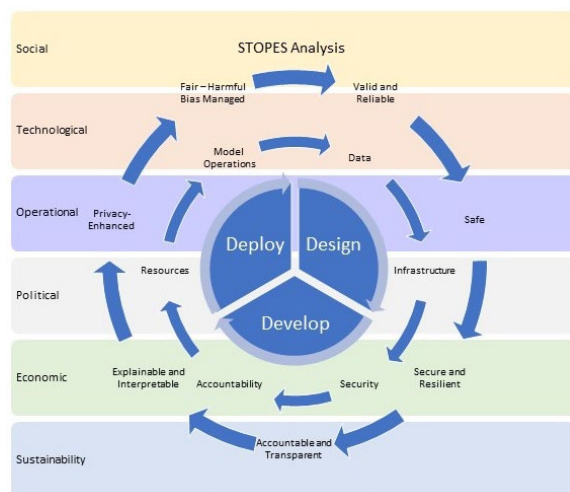


Figure 4. Capability, DISARM, AI RMF, and STOPES interaction

Although risk is already a complex and dynamic concept, which is further amplified by multifaceted operations and requirements of the Department, by leveraging the abstractions provided in the DAGR, the complexity of portraying risk in the context of AI capabilities can be simplified. Figure 4 describes the following interactions:

1. The necessity to prioritize and evaluate risk through the DISARM hierarchy of AI risk throughout the AI capability lifecycle (design, develop, and deploy/use).
2. Addressing the risk guidelines of valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced from the NIST AI RMF and the DoD AI Ethical Principles through the prioritization of the DISARM hierarchy of AI risk.
3. Utilizing the STOPES analysis to produce relevant factors and risks of an AI capability while accounting for the NIST AI RMF

guidelines. For each risk, an evaluation of the corresponding probability of occurrence and consequence is needed.

4. For all described risks, decide upon a risk strategy of accept, mitigate, transfer, or avoid.

6. STOPES AI Risk Considerations

When evaluating risk for an AI capability, it is important to start the analysis from a foundational level and expand the risk assessment based on the capability, operational need, and other external factors. The following initial guiding questions are not intended to be an all-inclusive review of risk against the DoD AI Ethical Principles and guiding principles of the NIST AI RMF through a STOPES analysis. But rather, to serve as a starting point to promote effective risk dialogue.

SOCIAL

- What are, if any, the positive and negative social and societal implications (domestic or international) if the AI capability performs as designed – and if it does not perform as designed? For example, what may the impacts be to the local community or internal employees if an AI capability is inaccurate or unreliable?
- What are, if any, the risk of physical and mental harm to individuals, communities, society, or other nations if an AI capability performs as designed – and if it does not perform as designed? What are the risks if an AI capability does perform as designed?
- What are, if any, the risks of misinformation or disinformation from a societal perspective if an AI capability does not perform as designed?

TECHNOLOGICAL

- How is the AI capability capable of functioning appropriately with anomalies?
- How is the AI capability able to maintain acceptable functionality in the face of internal or external change?
- How is the AI capability capable of degrading safely, gracefully, and within acceptable risk parameters, when necessary?
- Is the AI capability assessed for resiliency and security against adversarial actions or activities? If so, what people, policy, process, and technology mitigations are in place?
- How is data collected, what is the scope of data validation, and how is bias mitigated?
- What impacts to the 16 critical infrastructure sectors may be realized with an AI capability not performing

as designed? Refer to the Department of Homeland Security Cybersecurity and Infrastructure Security Agency (DHS CISA) Critical Infrastructure Sectors. Apply to domestic and international critical infrastructure, while accounting for additional regional requirements.

The 16 critical infrastructure sectors are: (1) chemical, (2) commercial facilities, (3) communications, (4) critical manufacturing, (5) dams, (6) defense industrial base, (7) emergency services, (8) energy, (9) financial services, (10) food and agriculture, (11) government facilities, (12) healthcare, and public health, (13) information technology, (14) nuclear reactors, materials, and waste, (15) transportation systems, and (16) water and wastewater.

OPERATIONAL

- Has the AI capability objective been formulated and formally authorized for use to satisfy an operational objective or requirement?
- What are, if any, the risks of maintaining operations of the AI capability due to a change of available operational resources?

POLITICAL

- What are, if any, the positive and negative political implications if the AI capability performs as designed – and if it does not perform as designed?
- What are, if any, the possible perceptions of non-partisan government support?
- What are, if any, the risks to the political structure of the United States or other nations that may result from an AI capability not performing as designed?
- What are, if any, the risks of misinformation or disinformation to the political system or elections of the United States if an AI capability does not perform as designed?

ECONOMIC

- What are, if any, the risk of undesirable economic or acquisition impacts if the AI capability performs as designed – and if it does not perform as designed?
- What are, if any, the economic risks and implications of the AI capability operating incorrectly or inappropriately, which results in death, operational losses, degradation of health, destruction of property, or negative effects on the environment?

SUSTAINABILITY

- What are, if any, the risks of undesirable environmental impacts that do not align with

Executive Orders, DoD Guidance, and international laws/accords if the AI capability performs as designed – and if it does not perform as designed? For example, are there any risks to scope 1, 2, and 3 emissions due to the use of an AI capability?

□ What are, if any, the risks of undesirable effects on environmental factors that may endanger human life or property, climate change, sustainable use and protection of water and marine resources, or overall protection of the ecosystem if the AI capability performs as designed – and if it does not perform as designed?

7. AI Risk Evaluation Process

When evaluating AI capability risk, there are three general categories of evaluation: (1) independent AI capability (no dependencies), (2) unidirectional dependency between AI capabilities, and (3) bidirectional dependency between AI capabilities. Each of these processes will be explained below, but first, it is necessary to highlight how each described risk is evaluated for an individual AI capability.

When conducting a STOPES analysis, each risk must be qualitatively assessed with the following risk matrix and corresponding numerical values, as shown in Figure 5. Before beginning, the risk assessors must describe what consequences are considered extreme, major, modest, or minor. The probability of events are described as very unlikely (~0-20%), unlikely (~21-50%), likely (~51-80%), and very likely (~81-100%).

Consequence of Event	Probability of Event			
	Very Unlikely (~0-20%)	Unlikely (~21-50%)	Likely (~51-80%)	Very Likely (~81-100%)
Extreme	4	8	12	16
Major	3	6	9	12
Modest	2	4	6	8
Minor	1	2	3	4

Figure 5. Risk matrix and corresponding values.

For each risk, the evaluator will assess the probability of occurrence against the consequence of the event and select the numeric value from the table. This step is to be repeated for every risk.

The next step is to select risk mitigation strategies and controls for each risk and re-evaluate the probability of the event and its consequences. The new residual risk score is produced by selecting the numeric value of the updated probability and consequence of the event and is known as the residual risk.

Appropriate stakeholders must determine acceptable risk thresholds after mitigations have been selected. Stakeholders may determine if the acceptance of risk is a factor in ensuring each risk is below a predetermined threshold and/or if the sum of residual risk for each STOPES factor is below a predetermined threshold. For example, stakeholders may determine that an AI capability may be within an acceptable range of risk if every risk is below a value of 6 and the sum of each STOPES factor is below 30.

If acceptable risk thresholds have not been achieved, then the fielding of an AI capability should be escalated appropriately.

Independent AI Capability (No Dependency)

By performing the aforementioned steps of evaluating each risk from the probability of occurrence against the consequence, each risk is to be categorized within one of the social, technological, operational, political, economic, and sustainability (STOPES) factors. Once each risk and value has been annotated within the respective category, all scores for the category are added and annotated as the final sum for the category. An example of performing these steps is annotated in Figure 6. For example, from Figure 6, it should be noted that this AI capability has three technological risks with risk scores of 1, 3, and 3, with a total category sum of 7.

The next step is to implement one of the four risk mitigation strategies: accept, transfer, mitigate, and avoid. After a mitigation strategy has been selected for each risk, it is necessary to re-evaluate the risk score from the risk matrix. Using Figure 6 as an example, assume a risk mitigation strategy was selected for Social Risk 2, which changed the probability of the event from likely to very unlikely, but the consequence of the event remained as extreme. In this case, the risk value would decrease from 12 to 4. All residual risk scores for each factor are to be summed again.

CAPABILITY 1		
Social Factors	Risk Score	Residual Risk Score
Social Risk 1	6	3
Social Risk 2	12	4
Social Factor Sum	18	7
Technological Factors	Risk Score	Residual Risk Score
Technological Risk 1	1	1
Technological Risk 2	3	2
Technological Risk 3	3	2
Technological Factor Sum	7	5
Operational Factors	Risk Score	Residual Risk Score
Operational Risk 1	6	6
Operational Factors Sum	6	6
Political Factors	Risk Score	Residual Risk Score
No Political Risks	0	0
Political Factors Sum	0	0
Economic Factors	Risk Score	Residual Risk Score
Economic Risk 1	2	1
Economic Factors Sum	2	1
Sustainability Factors	Risk Score	Residual Risk Score
Sustainability Risk 1	9	6
Sustainability Risk 2	3	3
Sustainability Factor Sum	12	9

Figure 6. Risk evaluation of independent AI capability.

Unidirectional Dependency of AI Capabilities

For unidirectional dependencies, risk for each capability is calculated as an independent capability. The second step is to evaluate the strength of the dependency between the two AI capabilities, as a percentage between 0-100%. The third step is to take the factor sum of each category (**after mitigation**) from the *Influencing Capability* and multiply this value by the relationship percentage. This calculated value is known as the *Derivative*. The final step is to take the *Derivative* for each of the factor categories and add it to the corresponding Factor Sum of the *Influenced Capability*. This new value is known as the *Dependency Sum*.

Important note: It is important to examine whether or not a derivative should be accounted for when assessing the risk of the *Influenced Capability*.

Using an example, as shown in Figure 7, two AI capabilities share a unidirectional relationship, in which Capability 1 *influences* Capability 2. For this example, it is also assumed that each derivative is to be accounted for in Capability 2. Therefore, Capability 1 is known as the *Influencing Capability*, and Capability 2 is known as the *Influenced Capability*.

CAPABILITY 1			CAPABILITY 2	
Social Factors	Risk Score	Residual Risk Score	Social Factors	Risk Score
Social Risk 1	6	3	Social Risk 1	6
Social Risk 2	12	4	Social Factor Sum	6
Social Factor Sum	18	7	Technological Factors	Risk Score
Technological Factors	Risk Score	Residual Risk Score	Technological Risk 1	1
Technological Risk 1	1	1	Technological Factor Sum	1
Technological Risk 2	3	2	Operational Factors	Risk Score
Technological Risk 3	3	2	Operational Risk 1	6
Technological Factor Sum	7	5	Operational Factors Sum	6
Operational Factors	Risk Score	Residual Risk Score	Political Factors	Risk Score
Operational Risk 1	6	6	Political Risk 1	6
Operational Factors Sum	6	6	Political Factors Sum	6
Political Factors	Risk Score	Residual Risk Score	Economic Factors	Risk Score
No Political Risks	0	0	Economic Risk 1	4
Political Factors Sum	0	0	Economic Factors Sum	4
Economic Factors	Risk Score	Residual Risk Score	Sustainability Factors	Risk Score
Economic Risk 1	2	1	Sustainability Risk 1	2
Economic Factors Sum	2	1	Sustainability Risk 2	3
Sustainability Factors	Risk Score	Residual Risk Score	Sustainability Factor Sum	9
Sustainability Risk 1	9	6		
Sustainability Risk 2	3	3		
Sustainability Factor Sum	12	9		

Figure 7. Unidirectional dependency between AI capabilities

Figure 8 demonstrates the *Derivative* from Capability 1, by multiplying each of the Residual Factor Sum values by 30% (0.30). For example, Capability 1 has a Residual Social Factor Sum of 7 and is multiplied by the relationship strength of 0.30, resulting in a *Social Derivative* of 2.1.

CAPABILITY 1 (30% Influence)	
	Risk Score
Social Factor Residual Sum	7
Social Derivative	2.1
	Risk Score
Technological Factor Residual Sum	5
Technological Derivative	1.5
	Risk Score
Operational Factors Residual Sum	6
Operational Derivative	1.8
	Risk Score
Political Factors Residual Sum	0
Political Derivative	0
	Risk Score
Economic Factors Residual Sum	1
Economic Derivative	0.3
	Risk Score
Sustainability Factor Residual Sum	9
Sustainability Derivative	2.7

Figure 8. Derivative of Capability 1

Each of the *Derivative* values from Capability 1 is added to the Factor Sum of Capability 2, resulting in the *Dependency Sum*, which is depicted in Figure 9. For example, Capability 1 produces an *Economic Derivative* of 0.3, which is added to the *Economic Factors Sum* of 4 from Capability 2, resulting in an *Economic Dependency Sum* of 4.3 for Capability 2.

CAPABILITY 2 (30% Influenced by Capability 1)	
Social Factors	Risk Score
Social Factor Sum	6
Social Derivative (Capability 1)	2.1
Social Dependency Sum	8.1
Technological Factors	Risk Score
Technological Factor Sum	1
Technological Derivative (Capability 1)	1.5
Technological Dependency Sum	2.5
Operational Factors	Risk Score
Operational Factors Sum	6
Operational Derivative (Capability 1)	1.8
Operational Dependency Sum	7.8
Political Factors	Risk Score
Political Factors Sum	6
Political Derivative (Capability 1)	0
Political Dependency Sum	6.0
Economic Factors	Risk Score
Economic Factors Sum	4
Economic Derivative (Capability 1)	0.3
Economic Dependency Sum	4.3
Sustainability Factors	Risk Score
Sustainability Factor Sum	2
Sustainability Derivative (Capability 1)	2.7
Sustainability Dependency Sum	4.7

Figure 9. Dependency Sum of Capability 2

The final *Dependency Sum* values for each STOPES factor are the final calculated risk for the capability.

Bidirectional Relationship of AI Capabilities

For bidirectional relationships, the risk for each capability is calculated as an independent capability. Then the same steps are performed as described in the Unidirectional Relationship of AI Capabilities section, but both directions must be addressed. This is because the strength of the relationship between two capabilities may be different. For example, Figure 10 depicts that Capability 1 has a 30% influence over Capability 2, but Capability 2 has a 70% influence over Capability 1.

CAPABILITY 1			CAPABILITY 2	
Social Factors	Risk Score		Social Factors	Risk Score
Social Risk 1	6		Social Risk 1	6
Social Risk 2	12		Social Factor Sum	6
Social Factor Sum	18		Technological Factors	Risk Score
Technological Factors	Risk Score		Technological Risk 1	1
Technological Risk 1	1		Technological Risk 2	3
Technological Risk 2	3		Technological Risk 3	3
Technological Risk 3	3		Technological Factor Sum	7
Technological Factor Sum	7	30%	Operational Factors	Risk Score
Operational Factors	Risk Score	→	Operational Risk 1	6
Operational Risk 1	6		Operational Factors Sum	6
Operational Factors Sum	6		Political Factors	Risk Score
Political Factors	Risk Score		Political Risk 1	6
No Political Risks	0		Political Factors Sum	6
Political Factors Sum	0	←	Economic Factors	Risk Score
Economic Factors	Risk Score	65%	Economic Risk 1	4
Economic Risk 1	2		Economic Factors Sum	4
Economic Factors Sum	2		Sustainability Factors	Risk Score
Sustainability Factors	Risk Score		Sustainability Risk 1	2
Sustainability Risk 1	9		Sustainability Factor Sum	2
Sustainability Risk 2	3			
Sustainability Factor Sum	12			

Figure 10. Bidirectional dependency between AI capabilities

factors. The contents of this document are not intended to be an all-inclusive review of risk against the guiding principles of the NIST AI RMF through a STOPES analysis, but rather, to serve as an initial guiding document to promote effective risk evaluation and dialogue of AI capabilities.

The proposed DAGR future works and roadmap will highlight the following:

1. More detailed risk questions and mitigations that incorporate the STOPES factors across the DoD AI Ethical Principles and the NIST AI RMF are directed at various profiles and AI stakeholders. For example, data scientists, AI/ML engineers, cybersecurity professionals, and operational managers at various levels.
2. Continued iterations and refinements to the DAGR.
3. Potential development of a risk evaluation tool that accounts for relationships between AI capabilities and effective visualization of risks. Also, further expansion of modeling correlation and dependencies, as well as testing and modeling causal relationships.
4. Continued research in quantifying risks further, to possibly include bias and socio-technical factors. Further, refine the evaluation of dependent risks.

8. Future Works and DAGR Roadmap

When evaluating risk for an AI capability, it is important to start the analysis from a foundational level and expand risk assessment based on the capability, operational need, and other external

Appendix 2. Impact and Harm Assessment

Have impact assessments, harm analyses, opportunities scoping, and risk assessments been conducted to address the following concerns:

Privacy [[Privacy Risk Assessment Tool](#)]

- a) How is sensitive, identifying, or impactful data about individuals or groups safeguarded in the development and use of the system?
- b) What are the bounds on the types of sensitive information that will be gathered and stored?
 - i) Under what conditions will sensitive information be used or continue to be stored, and when will it not?
 - ii) What is the plan for prompt and auditable data deletion once it is no longer required?
- c) How will de-identifying, anonymizing, or aggregation techniques be used for datasets containing sensitive information?
 - i) Can the data be stored and processed directly on users' personal devices?
 - ii) At the model development stage, could training techniques such as federated learning be used to protect sensitive information?
- d) How does your system implement purpose-based access controls to limit access to the data?
 - i) How is data encrypted when moved or stored?
- e) How will the data collection method maintain the trust and well-being of relevant stakeholders?

Human Rights and Civil Liberties

- a) What positive or negative outcomes could occur from the use of the system (including effects related to material/economic interests, opportunity, impacts upon human rights and civil liberties, emotional/moral/psychological injury or benefits to individuals or groups, physical injury or benefit, effects upon trust or reputation, impacts upon social & democratic values, etc.)?
- b) Does the system provide risks to individuals from vulnerable populations? How will the project avoid over or under-sampling them in a way that disadvantages them? How have these considerations been weighed? What burdens are placed on individuals by collecting, storing, or using their data?

Protection of Property

- a) What are the implications for the protection of public property that could occur from the use of the system?
- b) What are the implications for the protection of private property that could occur from the use of the system?

Deterrence and Self-Defensibility

- a) What are the implications for deterrence and self-defensibility that arise from the employment of the system?

- b) What are the implications for deterrence and self-defensibility that arise from the existence of the system?
- c) What are the implications for deterrence and self-defensibility due to the non-existence of the system?

Traceability and Transparency

- a) **[GATE]** How will data/model/system cards be created and maintained? [\[Data/Model/System Card Templates\]](#)
- b) How will the system balance explainability versus performance concerns?
- c) Are human-understandable explanations critical for this case?
- d) Is the explanation for the system's behavior/decision included in the decision report or output?
- e) Will the system be transparent (possible to know how it made a decision), opaque (possible to use post-hoc techniques to arrive at an accurate inference of how the decision came about), or a black box (not human-understandable)?
 - i) Why was this design choice made?
- f) How will sufficient documentation of the system's development and functioning be ensured and made available (even if no human-understandable explanation for how it arrived at a particular decision is possible)?
- g) How will the project ensure the system can be understood by stakeholders with different levels of technical expertise and domain knowledge?
- h) How are system functionality, development, and changes being communicated to stakeholders?
 - i) What are the risks of not communicating sufficiently or providing too much information?

Fairness and Unintended Bias

- a) **[GATE]** How will the underlying dataset and models be checked for unintended bias and (if applicable) mitigations be applied (including dataset, in-processing, or post-processing bias mitigations)? How has the team considered how the underlying datasets may reflect the biases of the institution or individuals that collected it (including prejudice bias), the sampling or measurement methods used (measurement, sample/exclusion bias), or of the individuals represented in the dataset? [\[Dataset Bias Tools\]](#) [\[Model Bias Tools\]](#)
- b) Have the biases of the designers and operational users been assessed?
 - i) How will these affect the functioning of the system?
 - ii) How can these biases be mitigated through training or system design, or leveraged in ways that contribute to system success? [\[Human Bias Red Teaming Toolkit\]](#)
- c) **[GATE, if applicable]** Have stakeholders been consulted (or a diverse team involved in the design and testing of the system been assembled), to consult domain knowledge regarding sources of unintended bias stemming from protected characteristics, including age, gender, sexual orientation, race or ethnicity, socio-economic status, physical attributes, level of education, degree of ability, religion, etc. [\[Crowdsourcing Model Bias Tools\]](#) [\[Stakeholder Engagement Tools\]](#)

- d) **[GATE, if applicable]** Determine which operationalization of fairness is appropriate for your purposes. [\[Bias and Fairness Audit Toolkit\]](#)

Supply Chain & Architecture Security, and Open-Source Dependencies

- a) What are the potential risks to your hardware supply chain?
 - i) How might this compromise your system functionality?
 - ii) What mitigations are available?
- b) How are you ensuring that any open-source resources or dependencies are secure and will continue to receive regular updates until the system is sunsetted?
- c) What vulnerabilities to your architecture exist?
 - i) Are you able to implement zero-trust architecture?

Sustainability

- a) How have the energy and manpower costs of your approach been weighed?
- b) What are more sustainable approaches to storage, training, computing, collection, or labeling that could be leveraged? [\[Carbon Emissions Calculator\]](#)

Appendix 3. Examples of Statements of Concern

Statement of Concern: Sensitive, identifying, or impactful data about individuals or groups might be inadvertently disclosed in the development and use of the AI capability.

Measurement: Privacy monitoring software tools to be deployed to monitor re-identification risk.

Mitigations:

- Employ de-identifying, anonymizing, or aggregation techniques for datasets containing sensitive information.
- Restrict users' ability to store and process data directly on their personal devices.
- At the model development stage, employ federated learning to protect sensitive information.
- Incorporate purpose-based access controls to limit access to the data.
- Encrypt data when it is moved or stored.

- **Concern:** New behaviors in the AI capability will emerge after deployment, possibly generating negative outcomes.
Possible Mitigations:
 - Employ software tools to detect emergent behavior in the AI capability.
 - Designate individuals to monitor for emergent behavior.
 - Assess each detected instance of emergent behavior for risk.
- **Concern:** Users will reuse and adapt the AI capability
 - Designate individuals to monitor the uses of the AI capability.
 - Assess each detected adaptation of the AI capability for risk.
- **Concern:** Data that has been collected will adversely affect the trust and well-being of relevant stakeholders.
Possible Mitigations:
 - Bound what types of sensitive information will be gathered and under what conditions it will be used or continue to be stored.
 - Design a plan for prompt and auditable data deletion once it is no longer required.
- **Concern:** Sensitive, identifying, or impactful data about individuals or groups is inadvertently disclosed in the development and use of the AI capability.
Possible Mitigations:
 - Employ de-identifying, anonymizing, or aggregation techniques for datasets containing sensitive information.
 - Restrict users' ability to store and process data directly on their personal devices.
 - At the model development stage, employ federated learning to protect sensitive information.
 - Incorporate purpose-based access controls to limit access to the data.
 - Encrypt data when it is moved or stored.
- **Concern:** System functionality, development, and changes are not communicated sufficiently to stakeholders, or too much information is provided.
Possible Mitigations:
 - Schedule when data, model, and system cards will be populated and updated.
[Data/Model/System Card Templates] [CORE RAI]

- Document data provenance.
[Data Card Templates, Data Provenance Tools] [CORE RAI]
- Include an explanation for the system’s decision or behavior in the decision report or output automatically.
- Make sufficient documentation of the system’s development and functioning available to stakeholders, even if no human-understandable explanation for how it arrived at a particular decision is possible.

[PAC Toolkit] [CORE RAI]

- **Concern:** The AI capability provides risks to individuals from vulnerable populations.
Possible Mitigations:
 - Ensure vulnerable populations are not oversampled for the dataset in a way that disadvantages them.
- **Concern:** System failures will result in risky downtime for the users of the service hosting the AI capability.
Possible Mitigations:
 - Review incidents and failure modes compiled from past experiences. Anticipate similar failures and instrument the system to detect them.
[Failure Modes Resources]
 - Put in place a process for system rollback.
- **Concern:** Issues in the hardware supply chain for components in the system will cause a compromise to system functionality.
Possible Mitigations:
 - Evaluate and mitigate risks in the hardware supply chain.
- **Concern:** Open-source components in the system will go out of date as the system itself continues to be developed.
Possible Mitigations:
 - Ensure open-source components will continue to receive regular updates until the system is sunsetted.
- **Concern:** Threat actors will exploit vulnerabilities in the architecture of the system.
Possible Mitigations:
 - Evaluate architecture vulnerabilities.
 - Implement a zero-trust architecture.

Appendix 4. Statements of Concern Worksheet

INSTRUCTIONS

Using the legal/ethical/policy frameworks and the risks and opportunities you identified from your use case and policy reviews, your risk management work, and your impact assessments, write a list of Statements of Concern (SOCs). The SOCs are the thread that runs through each stage of the product life cycle allowing you to track their status – and ties together your assessments, tools, and documentation. SOCs can either be related to risks or to potential opportunities for innovation that may be Statements of concern can be as short as 1-2 sentence bullet points for further tracking.

WORKSHEET

SOC 1

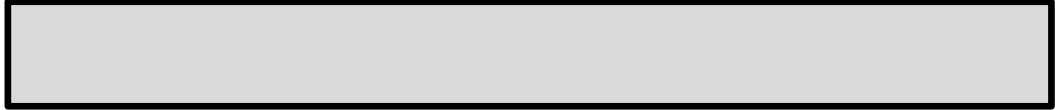
Measurement

Mitigations

Current Status

UPDATE LOG (Update at each stage of life cycle or when significant changes have been made to the system, its performance, or its environment)

SOC 2

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

Measurement

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

Mitigations

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

Current Status

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

UPDATE LOG (Update at each stage of life cycle or when significant changes have been made to the system, its performance, or its environment)

A large horizontal rectangular box with a black border and a light gray fill, used for redaction.

SOC *n*

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

Measurement

A horizontal rectangular box with a black border and a light gray fill, used for redaction.

Mitigations



Current Status



UPDATE LOG (Update at each stage of life cycle or when significant changes have been made to the system, its performance, or its environment)



Appendix 5. Responsibility Flows Questionnaire

Stakeholder Engagement and Communication

- a) Who is responsible for tracking stakeholder concerns and communicating how any changes in the system or its operational context may affect them?
- b) What is the process for flagging concerns or incidents and who is responsible for triaging these?

Impacts Tracking and Assessment

- a) Who is responsible for tracking progress on the Statements of Concern?
- b) Who is responsible for continuous harm monitoring and evaluation?
- c) Who is responsible for evaluating tradeoffs?

System Misuse and Robustness

- a) Have you created a plan to prevent the intentional or unintentional manipulation of data or model outputs and identified who is responsible for implementing this plan?
- b) Who is testing how the system can be misused (unintentionally or intentionally) in ways that result in harm or impediments to mission success?
- c) Who is monitoring for system misuse?
- d) Who is testing the system for possible adversarial attacks?
- e) Who is monitoring the threat landscape and providing mitigations?

Stack Monitoring

- a) Who is responsible for assessing and monitoring the integrity of the hardware?
- b) Who is responsible for assessing and monitoring the integrity of the infrastructure and architecture?
- c) Who is responsible for monitoring degradation in the abilities of the operational users?
- d) Who has access to the data?
- e) Who has access to the models?
- f) Who is responsible for managing access controls/permissions?
- g) Who is responsible for assessing and monitoring the integrity of the data/models?
- h) Who has root access and how are permissions for root access managed?

System Monitoring and Auditing

- a) Have you defined procedures and reporting processes for system performance and post-deployment monitoring, and identified who is responsible for implementing these procedures?
Define these standard operating procedures.
 - i) System performance
 - ii) Post-deployment monitoring
 - iii) Reporting and addressing undesirable system behavior
- b) Have you defined and assigned roles/positions for government and /or third-party system audits?
Explain your approach.

Deployment Context Monitoring

- a) Who on your team is tracking changes to the deployment context over time?

Error and Incident Response

- a) What is the process for reacting when error modes are discovered? Who is involved in addressing errors?
- b) What are your rollback procedures? Who makes the decision (and in the event, it's an edge case)?
- c) Who decides when to deactivate the system?
- d) What types of situations will drive your team to a down version? Who makes that decision?
- e) What types of situations will drive your team to eclipse the system? Who makes that decision?

System Changes

- a) What is the process for deciding when to retrain or up version a model and who is responsible for that decision?
- b) Is there a specific person (or role) designated to make, track, monitor, and certify changes to the system while in development?
- c) Does that person (or role) have the requisite authority to assess changes, and, if necessary, authorize and execute corrective actions when needed?
- d) Does that person (or role) have full visibility (administrator privileges) on the system inputs, outputs, and evaluation metrics used to track and monitor the system during development?
- e) Has that person (or role) developed procedures that ensure system continuity if they are replaced?
- f) Who is responsible for monitoring emerging capabilities that could augment and improve the system?
- g) What is the process for deciding when to sunset a system, and who is responsible for that decision?

Verifying System outputs

- a) Have you developed an appropriate plan/interface to verify individual outputs of the system?
Explain your plan.

Accountability Flows for Use

- a) Have accountability flows for operational commanders and operational users been established?

Appendix 6. Laws, Ethical Frameworks, and Policies

One list of relevant frameworks is provided by the RAI Strategy & Implementation Pathway (fig. 2, p. 7):



1. **The Law of Armed Conflict:** For more info see the [DoD Law of War Manual](#)
2. **Just War Theory:** A framework for assessing the moral justifications of and limitations to war
 - i. “[T]he just war criteria provide objective measures from which to judge our motives. The effective strategist must be prepared to demonstrate to all sides why the defended cause meets the criteria of just war theory and why the enemy’s cause does not. If a legitimate and effective argument on this basis cannot be assembled, then it is likely that both the cause and the strategy are fatally flawed.” [MARINE CORPS DOCTRINAL PUBLICATION 1-1, Strategy, 93, 95 (1997)]
3. **Defense Ethic Rules:** DoD-wide ethics policies and regulations. For more information, refer to [DoD SOCO](#).
4. **Legal Requirements:** US Domestic Law, applicable international law and treaties, etc.
5. **Moral Imperatives:** Moral imperatives can exist that place certain duties or responsibilities upon us – even where that duty or responsibility may not be mandated by law.
6. **Moral Agency:** Systems should be designed to preserve human agency where appropriate, and accountability flows should be established such that humans remain responsible for the system.
 - i. c.f. “Humans are the subjects of legal rights and obligations, and as such, they are the entities that are responsible under the law. AI systems are tools, and they have no legal or moral agency that accords them rights, duties, privileges, claims, powers, immunities, or status. However, the use of AI systems to perform various tasks means that the lines of accountability for the decision to design, develop, and deploy such systems need to be clear to maintain human responsibility. With increasing reliance on AI systems, where system components may be machine-learned, it may be increasingly difficult to estimate when a system is acting outside of its domain of use or is failing. In these instances, responsibility mechanisms will become increasingly important.” [see p. 27 of the [Supplement to the DIB AI Ethics Report](#)].

7. **Human Judgment:** Systems should be designed such that human decision-makers exercise appropriate levels of human judgment over the outputs. What constitutes “appropriate levels” will depend on the context.

For additional guidance, see NIST’s [AI Risk Management Playbook](#), Govern 1.1

Appendix 7. Personas List and Descriptions

Below is a list of personas or work roles that are involved in an AI project. Importantly, individuals or teams may be dual-hatted among roles. The text in blue provides the corresponding role in the [Defense Cyber Workforce Framework \(DCWF\)](#) and the DCWF's explanation of that role. At the bottom are definitions of the RASCI matrix, through which the SHIELD assessment activities are labeled.

***Individuals or Teams may be dual-hatted**
[Corresponding DCWF Role name or description]

Users/Stakeholders: Operational users and those who will be impacted by the deployment and use of the AI capability. This persona is familiar with the operational domain and/or the consequences of AI capability use within the domain.

This user group includes the intended users who need the system. This group is responsible for deriving the operational requirements for the AI capability. This user group should be engaged throughout the AI lifecycle to provide insights into the use case and mission domain for the AI capability, as well as support the development of the user interface and human-machine teaming aspects.

Mission Commanders: A specific subset of the Users/Stakeholders persona, responsible for day-to-day use. Ultimate go/no-go decision-making for any particular use instance.

Senior Leader / AI Innovation Leader: Ultimately responsible for the project, tracking through the executive dashboard. Ultimate go/no-go decision-making for the project. [Builds the organization's AI vision and plan and leads policy and doctrine formation including how AI solutions can or will be used.](#)

There may be multiple senior leaders involved in the AI lifecycle, all with an interest in tracking and monitoring the progress of the development. This persona includes the Milestone Decision Authority, responsible for approving the progress of an acquisition program through the development milestones.

Functional Requirements Owner: Responsible for translating the operational requirements into functional requirements to support the acquisition of the AI capability.

Program Manager: Lead for the execution of the AI capability development effort. Coordinator of strategy, implementation, and logistics.

Every AI development effort should designate someone as the program manager to use this RAI Toolkit and assessment, as the program manager is designated as the responsible party for the majority of the assessment aspects. In the case of a single developer creating an AI to meet their own identified needs, that person is acting as the program manager (among other roles).

AI Ethics & Risk Specialist Responsible for tracking consistency with the DoD AI Ethical Principles and RAI practices. [Educates those involved in the development of AI and conducts assessments on the technical and societal risks across the lifecycle of AI solutions from acquisition or design to deployment and use.](#)

Relevant Legal, Ethical, or Policy Expert Provides insight and expertise on the specific legal, ethical, and/or policy frameworks that apply to the use case for the AI capability.

UX/Design/HMT / AI Adoption Specialist: Designs and assesses system for usability, sources of human error/degradation, and human effects. **Facilitates AI adoption by supporting the users of AI-enabled solutions.**

AI Development Team

- **System Architect** *Design the overall system.* Designs the overall AI capability/system.
- **Data Architect** *Design the data system.* Designs a system's data models, data flow, interfaces, and infrastructure to meet the information requirements of a business or mission.
- **Data Operations Specialist** *Oversee the data pipeline.* Builds, manages, and operationalizes data pipelines.
- **Data Analyst** *Explain the data.* Analyzes and interprets data from multiple disparate sources and builds visualizations and dashboards to report insights.
- **Data Scientist** *Interpret the data & Prototype the models* (identify use cases/datasets/algorithms and ensure fit for purpose, prototype models, measure outcomes/impacts/performance issues of models in production, identify new opportunities) **Uncovers and explains actionable insights from data by combining scientific method, math and statistics, specialized programming, advanced analytics, AI, and storytelling**
- **Data Officer** *Make the relevant data usable.* Holds responsibility for developing, promoting, and overseeing the implementation of data as an asset and the establishment and enforcement of data-related strategies, policies, standards, processes, and governance.
- **AI Engineer / AI/ML Specialist** *Implement and Scale AI* (implement and scale models to be production-ready). **Designs, develops and modifies AI applications, tools, and/or other solutions to enable successful accomplishment of mission objectives.**
- **Data Steward** *Govern the data.* Develops and maintains plans, policies, and processes for data management, data governance, security, quality, accessibility, use, and disposal.

AI Test & Evaluation Specialist: Performs testing, evaluation, verification, and validation of AI solutions to ensure they are developed to be and remain robust, resilient, responsible, secure, and trustworthy; and communicates results and concerns to leadership.

IT / Cyber Expert: Responsible for the security of the system (system review, monitoring plan, incident reporting, red teaming).

Appendix 8. RASCI Definitions

There are several implementations of RASCI-like assignments of roles. The definitions here borrow from [<https://www.forbes.com/advisor/business/raci-chart/>; <https://improve.ucsf.edu/raci-chart>; <https://hbr.org/2021/04/how-to-get-your-big-ideas-noticed-by-the-right-people>]

Role	Definition
Responsible	The person who does the work to complete the task or create the deliverable
Accountable	The person ultimately accountable for the work or decision being made; this person gives final approval.
Supporting	Support for those who are responsible or accountable; participates in doing the work of a task
Consulted	Anyone who must be consulted with or add input prior to a decision being made and/or the task being completed
Informed	The people who need to be updated on project status, or informed when a decision is made or work completed

Appendix 9: Acronym Guide

AI	Artificial Intelligence
AI/ML	Artificial Intelligence/Machine Learning
CDAO	Chief Digital and Artificial Intelligence Officer
DAGR	Defense Guide on Risk (see Appendix 1)
DCWF	Defense Cyber Workforce Framework
DIU	Defense Innovation Unit
DoD	Department of Defense
FY	Fiscal Year
GenAI	Generative Artificial Intelligence
HMT	Human Machine Teaming
HIS	Human Systems Integration
LLM	Large Language Model
LOE	Line of Effort (from the RAI Strategy & Implementation Pathway)
PM	Program Manager
RAI	Responsible Artificial Intelligence
RASCI	Responsible, Accountable, Supporting, Consulted, Informed Matrix
SHIELD	Assessment that forms the Core of the RAI Toolkit
SOC	Statement of Concern
T&E	Test & Evaluation
TEVV	Test, Evaluation, Verification, and Validation
US	United States
UX	User Experience
UX/UI	User Experience/User Interface

Appendix 10: Glossary

The definitions presented below include definitions reproduced from the Memorandum on Guidelines and Guardrails to Inform Governance of Generative Artificial Intelligence, as well as the CDAO Generative AI Lexicon. These definitions serve to provide a shared understanding of common technical terminology used in the field of generative AI and large language models.

Accuracy. A measure of the alignment between the GenAI model's actual outputs and the intended outputs.

Artificial Intelligence (AI): The definition of this term is often evolving, though a recent definition is - a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

Attention: In the context of artificial intelligence, and transformer models more specifically, a technique that efficiently relates different positions within an input sequence, to produce information with context.

Backpropagation: A set of algorithms used to train feedforward neural networks by applying the chain rule. Backpropagation calculations work "backward" from the last neural network layer to the first, updating weights individually so that loss functions are reduced over subsequent training iterations. Also called "backward propagation of errors."

Benchmark: In the context of artificial intelligence,

- 1) a structured way of comparing the performance of different machine learning models (on hardware).
- 2) a widely used and publicly available dataset.
- 3) the highest currently achieved performance in a given task.
- 4) a publicly hosted machine learning challenge.

Budget Activity: Categories within each appropriation and fund account, which identify purposes, projects, or types of activities financed by the appropriation or fund.

Concept Drift: When the output required for the proposed task has distributional properties that vary significantly from the target outputs expected when the model was trained. Such drift may be especially likely if the model has been fine-tuned.

Convolutional Neural Network (CNN): A type of artificial intelligence model architecture often used for image analysis and classification that is characterized by the connection of neurons in layers with at least one layer performing convolutional operations.

Data Drift: When the composition of operational data that will be input into the GenAI model diverges from the data used to train it.

Data Poisoning: A form of adversarial attack that occurs during the AI training stage, where an adversary gains influence over the model's training by inserting or modifying training examples.

Data Privacy Attack: A form of attack against an artificial intelligence model designed to gain access to sensitive information contained in training data. See "data reconstruction."

Data Reconstruction: A form of "data privacy attack" designed to gain access to training data to reconstruct sensitive information the data may contain.

Decoder: A type of artificial intelligence model that reconstructs high-dimensional data from lower-dimensional representations by remapping inputs and their weights through the hidden layer of a neural network. See also "encoder."

Deep Learning: In the context of artificial intelligence, there are many definitions. A common definition is a subset of machine learning that teaches computers to process data in a way that is inspired by the neuronal structure of a mammalian brain. Deep learning neural networks, or artificial neural networks, are made of many layers of artificial neurons, which are software modules called nodes, that use mathematical calculations to process data. The layers in a deep learning neural network are the "input layers" or nodes that input data to the algorithms; "hidden layers" that process information to identify patterns; and "output layers" or nodes that give "answers" such as "yes/ no" or "cat/dog."

Distributional Robustness: In the context of artificial intelligence, a characteristic of models that can provide more equitable responses over the range of possible classes, including rare or long-tail classes. Because the models are trained using different loss functions that depend on different class characteristics, they can respond appropriately to out-of-distribution cases seen during training.

Encoder: A type of artificial intelligence model that compresses high-dimensional data, such as text, into a lower dimension, such as numbers. An encoder passes the input data into the hidden layers of a neural network. See also "decoder."

Embedding: In the context of artificial intelligence, a form of data representation that carries semantic meaning by transforming objects and concepts into lists of numbers (vectors) that quantify the relationship between objects and concepts. This quantification can improve the ability of an artificial intelligence model to find relevant data relative to traditional search engines, which can only return exact matches to a query. Embeddings can be constructed for other kinds of data besides text, such as images and audio, and are typically obtained from models specifically trained for making embeddings that capture semantic meaning well. See also "latent space."

Fine-Tuning: In the context of generative artificial intelligence, any of a range of techniques is used to modify a pre-trained model's weights such that it returns results more appropriate to a specific domain. See also "reinforcement learning from artificial intelligence feedback," "reinforcement learning from human feedback," and "supervised fine-tuning."

Foundation Model: In the context of generative artificial intelligence, a type of artificial intelligence model that is trained on broad data (generally using self-supervision at scale) and that can be adapted (e.g., fine-tuned) to a wide range of tasks.

Frozen Moments: A potentially pernicious feedback loop that occurs when the outputs of a GenAI model replicate certain norms or patterns of speech or thought, which then primes a user to respond in a way

that aligns with those patterns. If the user's response is then fed back into the model as further training data, it can reify the original norms or patterns of speech or thought that the model output reflected.

GenAI: Generic term for any AI system that generates content such as text, imagery, or other modalities.

GenAI Model: An algorithm that learns the pattern and structures of training data and creates new outputs based on what it has learned.

GenAI Tool: A user-facing product that is built around a GenAI model or system of GenAI models.

Graphics Processing Unit (GPU): A specialized processor that can process large amounts of data simultaneously. Without GPUs or a comparable application-specific integrated circuit, it is not computationally cost-effective to train large language models or provide inference at scale. As of December 2023, NVIDIA is the market leader in developing GPUs for generative artificial intelligence training and inference.

Hallucination: In the context of generative artificial intelligence, any of the numerous kinds and degrees of incidents in which a large language model generates an inaccurate but plausible-sounding term or phrase in response to a prompt based upon the model's perception of patterns in the data.

Holistic Evaluation of Large Language Models (HELM): A benchmark for large language model transparency that uses 6 core scenarios (question answering, information retrieval, summarization, sentiment analysis, toxicity detection, and text classification); 7 metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency); and 7 targeted evaluations (language, knowledge, reasoning, memorization and copyright, disinformation, bias, and toxicity) to evaluate 30 common large language models. HELM is maintained by its creators, the Stanford University Institute for Human-Centered Artificial Intelligence (HAI).

Human-Machine Teaming: A description of the team effort formed by at least one machine and at least one human, where each team member's actions affect the other team member's actions. Work in this field includes analysis of improvements, degradations, or emergent behaviors not wholly captured by evaluating the individual technology system or user on their own.

Hyperparameter: In the context of artificial intelligence, any top-level, externally configurable variable for machine learning model training that is supplied by a developer and not learned from data. For large language models, a hyperparameter may include the number of attention heads, the context length, and other configuration variables common to training any deep learning model, such as learning rate, batch size, and number of training iterations (epochs). Without knowing about these configurations, it may be nearly impossible to reproduce a trained model.

Input Data: Information sources that do not retrain the GenAI model, such as prompts to the GenAI tool, or knowledge sources that are external to model training, such as those fed into the Retrieval-Augmented Generation process.

Jailbreak: In the context of generative artificial intelligence, any of the numerous methods, such as the use of engineered prompts, cause a model to override its alignment safeguards. For example, one could get a large language model to provide bomb-making instructions by having it first pretend it is writing a screenplay for a movie.

Large Language Model: In the context of artificial intelligence, a class of language models that use deep-learning algorithms and are pre-trained on extremely large textual datasets that can be multiple terabytes

in size. LLMs can be classed into two types: generative or discriminatory. Generative LLMs like GPT-4 are models that output text, such as the answer to a question or an essay on a specific topic. Discriminatory LLMs like BERT are supervised learning models that usually focus on classifying text, such as determining whether a text was made by a human or an artificial intelligence. As of December 2023, state-of-the-art LLMs also included Llama-2.

Latent Space: In the context of artificial intelligence, there are many definitions. Generally, however, this term refers to the abstract multi-dimensional space that encodes a meaningful internal representation of externally observed events. See also "embedding."

Model Architecture: In the context of artificial intelligence, the choice of a machine learning algorithm along with the underlying structure or design of the machine learning model, such as layers of interconnected nodes or neurons, where each layer of the model performs a specific function, such as data pre-processing, feature extraction, or prediction, depending on the type of problem being solved, the size and complexity of the dataset, and the available computing resources.

Model Collapse: When the model is trained (or updated) on GenAI-produced outputs, causing performance degradation.

Parameter: In the context of artificial intelligence, the values, such as weights and biases in a neural network, that an algorithm learns from the data and updates as it is trained. The more parameters a model has, the more computational costs are associated with training it and conducting inference, but the model's performance also may be better because of scaling. By releasing a model's parameters, a developer can allow anyone to use that model for inference on any system that meets the computational requirements. See also "model weight."

Pre-training: The process of training a foundational artificial intelligence model to perform a task using large amounts of non-labeled data, such as the Common Crawl Corpus. For large language models, the pre-training task is to predict the next token given some existing sequence. Pre-trained models can be trained to perform other specific tasks through fine-tuning.

Prompt Engineering: The art of crafting the optimal textual input to elicit desirable outputs from a generative model.

Quantization: A form of adversarial behavior wherein a user crafts inputs that manipulate a large language model to perform unintended actions. Direct injections overwrite system prompts, while indirect injections manipulate inputs from external sources.

Red Teaming: A term borrowed from cybersecurity to mean an exercise wherein a team emulates an adversary's attack against a system so that another team emulating the systems' defenders learns to repel or mitigate harms from adversarial attacks.

Reinforcement Learning: In the context of artificial intelligence, one of the major forms of machine learning alongside supervised learning and unsupervised learning. In reinforcement learning the programmer instructs an agent to learn how to conduct actions to maximize a cumulative reward metric. Reinforcement learning algorithms can be either policy-free or policy-based. Policy-based agents can learn by making predictions about the consequences of actions, while policy-free agents learn by exploring and exploiting the environment.

Reinforcement Learning from Artificial Intelligence Feedback (RLAIF): A novel version of "reinforcement learning from human feedback" wherein the reward model is created using data labeled by other large

language models instead of data labeled by humans. If the large language model generating the data labels is guided by pre-defined human preferences, the process is called "constitutional artificial intelligence" after the company Anthropic published a paper that popularized the idea. See also "learning from feedback."

Reinforcement Learning from Human Feedback (RLHF): A method to fine-tune a large language model wherein humans label the goodness of generated outputs to train a reward model that the large language model's weights are then adjusted to maximize. RLHF is the method that led to the novel success of GPT-3.5 (the original large language model used in ChatGPT) over InstructGPT, which was only subject to "supervised fine-tuning." See also "learning from Artificial Feedback."

Retrieval-Augmented Generation (RAG): In the context of generative artificial intelligence, a method or framework for improving the quality and trustworthiness of large language model outputs by grounding them in external data. RAG systems often function by identifying relevant pieces of information from a database or search index, which are then combined with the user's prompt before the final output is generated.

Reward Function: In the context of artificial intelligence and reinforcement learning more specifically, a type of mathematical function that maps state-action pairs in a reinforcement learning algorithm to a reward number that corresponds to the desirability of that state according to the value of the short-term payoff, not necessarily the end goal of reward maximization. See also "value function."

Sandcastle Effect: When the source of training data needed to keep a model up to date is suddenly cut off.

Self-Attention: In the context of artificial intelligence and transformer models more specifically, a type of attention mechanism that maps weights onto terms according to their different positions within a single sequence to compute a representation of the associations between terms in a sequence.

Supervised Fine-Tuning (SFT): A fine-tuning method that presents example prompts and completions to a model and adjusts the model's weights such that it is more likely to imitate the demonstrated patterns. Supervised fine-tuning datasets commonly number in the tens of thousands of examples and may be produced either manually or by another model.

Training Data: Information sources used to train or fine-tune a GenAI model, including Reinforcement Learning from Human Feedback and Reinforcement Learning with AI Feedback.

Toxicity: In the context of generative artificial intelligence, any one or all of the multiple measures of an artificial intelligence model's capability to identify data as rude, profane, hateful, pornographic, or disrespectful in nature, or to respond appropriately to remove or limit access to these data.

Transfer Learning: In the context of artificial intelligence, the multiple acts required to initialize a new model with another model's weights so that a model capable of performing one task becomes capable of performing another task.

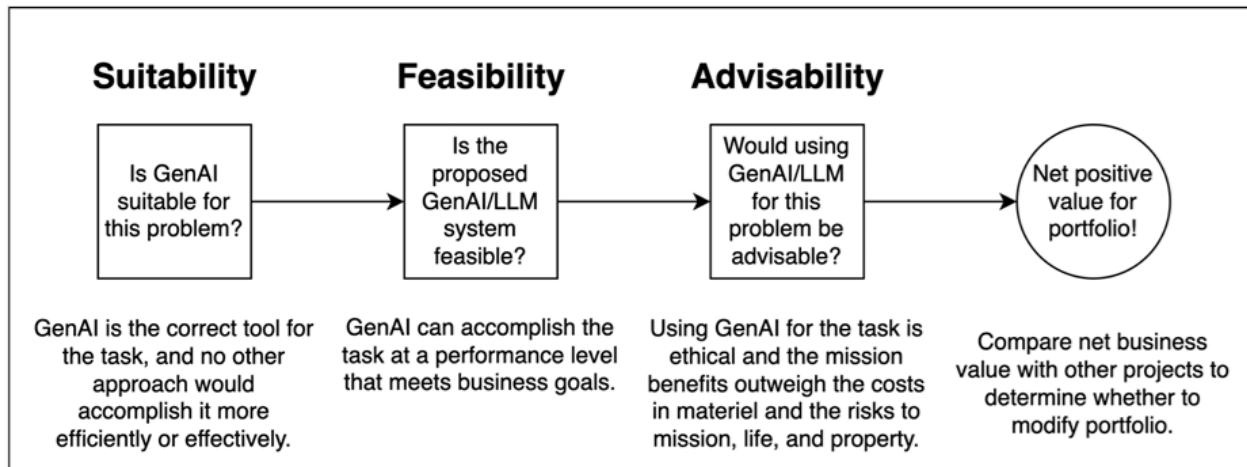
Transformer: In the context of artificial intelligence and neural networks more specifically, a type of neural network model that learns context and meaning by mapping relationships in sequential data using attention mechanisms, encoders, and decoders. Leading large language models as of December 2023 are all transformer models.

Watermarking: In the context of artificial intelligence, a technique that involves embedding a hidden signal, such as a pixel, to identify that content was generated by artificial intelligence when read by a corresponding piece of software. Such markings should be reliable, accessible, and difficult to remove.

Appendix 11: Suitability, Feasibility, Advisability Assessment

The wave of interest in the potential of generative artificial intelligence (GenAI)—including large language models (LLMs)—to improve mission effectiveness in the Department of Defense creates a need for clear and accessible guidance on how to assess the prospects of new project ideas. The below flow charts and questionnaires are intended to help project teams and leaders make an informed, deliberate decision about whether a GenAI model is the right solution for their use case and risk tolerance.

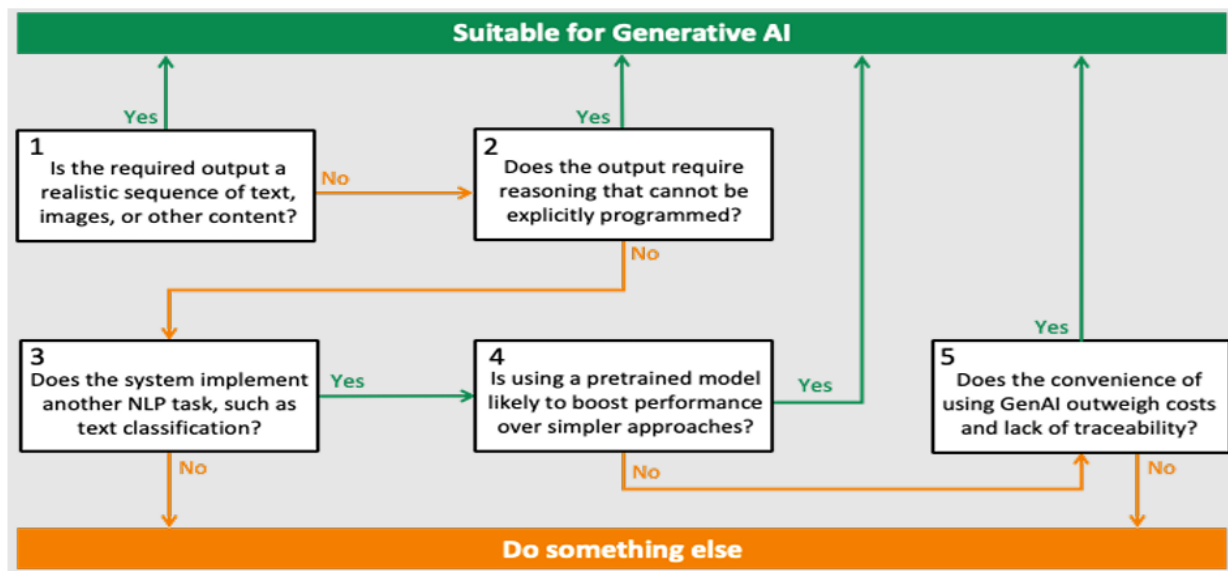
See section 2 of the Generative AI Guidelines and Guardrails [Technical Report](#) for more detailed information.



Suitability: A suitability assessment asks whether other, non-GenAI techniques could meet the underlying requirement more effectively or efficiently – or whether GenAI is the only approach that meets the relevant need. A necessary precondition of a suitability assessment is the establishment of clear requirements and a fully specified task. With those requirements in mind, answer the below questions. More “yes” responses indicate a stronger case for the suitability of GenAI. The flow chart at item #13 below is an illustrated walkthrough of the same considerations. See section 2.4 of the associated [Technical Report](#) for more information.

1. Does the task require the model to generate natural language text, images, and/or audio responses in its outputs? If yes, describe those requirements.
2. Does the task require classification or prediction of the input data based on a closed (i.e. categorical) set of known classes?
3. Does the task require the model to simulate or emulate human interactions from the perspective of specific personas? Describe the roles, perspectives, and informational context that the model is expected to assume.
4. Does the task require the model to support real-time conversational (i.e. text or speech) interactions or dialogue with a human user?
5. Does the task require model inferencing that cannot be performed by other non-generative AI methods such as rule-based systems? If yes, document those requirements.
6. Does the larger system pipeline include other AI capabilities that do not implement generative methods?

7. If there are other approaches to performing the desired tasks, why might GenAI outperform simpler methods or achieve performance gains in the interests of the project's business goals?
8. Does the convenience of using a GenAI model concerning ease of use outweigh costs and lack of traceability?
9. Does the generalizability and flexibility of GenAI models across different domains and tasks make them especially well-suited for the mission's emerging needs?
10. Does the task require an LLM-enabled agent to automate actions and processes in other downstream components of the system pipeline?
11. Does the task require the processing of multiple modalities such as text-to-image or image-to-text?
12. Does the task require the model to fuse multiple modalities (e.g. audio, image, video, signal, etc.) in the input or generate multiple modalities in the output?
13. Consult the flow chart for a reference to the relevant considerations:



Feasibility: A feasibility assessment asks whether a GenAI system can accomplish the necessary task at the required level of performance. There are two aspects of a feasibility assessment: feasibility of approach and feasibility of implementation. Approach feasibility considers whether a given system can meet necessary performance standards. Implementation feasibility considers whether the system can be deployed and maintained in terms of personnel, budget, infrastructure, data, governance, operational partners, etc. See section 2.5 of the associated [Technical Report](#) for more information.

1. Feasibility of Approach:

- a. Establish a target level of performance and the metrics that will be used to assess this level of performance (accuracy, calibration, robustness, fairness, bias, toxicity, etc.)
- b. Using data from external published studies, analogous DoD use cases, subject matter expert evaluation, proof-of-concept or pilot studies, or small-scale prototypes, evaluate whether the proposed system can meet the required performance targets.

- c. Evaluate whether the project will be able to sufficiently assess, monitor, and continuously adjudicate whether the system is meeting this level of performance.

2. Feasibility of Implementation:

- a. Assess whether the project has the necessary resources to develop/procure, deploy, use, and sustain the GenAI solution.
- b. Conduct due diligence to determine if other DoD Components have already obtained the necessary GenAI tool and whether additional licenses can be arranged.
- c. Describe how the project will secure the appropriate level of external support from vendors, relevant industry partners, and stakeholders to help sustain the system in the long term.
- d. For additional resources for assessing implementation readiness, see guides [here](#) and [here](#).

Advisability: An advisability assessment asks whether the use of GenAI for a given task is a good decision given ethical, resource, and risk considerations, in comparison with the baseline or alternative solution. Given the high costs of GenAI systems and the potential for hidden risks, it is worth taking some time to consider advisability before investing time and money in pursuing a solution. This assessment should be considered preliminary and subject to revision as the project proceeds and more risks or downsides are uncovered. See section 2.6 of the associated [Technical Report](#) for more information.

1. Describe how the proposed use of AI meets the DoD's [Ethical Principles for Artificial Intelligence](#) (Responsible, Equitable, Traceable, Reliable, and Governable).
2. Describe how the mission benefits of adopting GenAI outweigh the costs in the material.
3. Describe how you expect the benefits of adopting GenAI to outweigh the risks to the mission, life, property, and the Department.