Broad Agency Announcement

Autonomy Standards and Ideals with Military Operational Values (ASIMOV) Program

STRATEGIC TECHNOLOGY OFFICE

HR001124S0011

January 25, 2024

This publication constitutes a Broad Agency Announcement (BAA) as contemplated in Federal Acquisition Regulation (FAR) 6.102(d)(2) and 35.016 and 2 CFR § 200.203. Any resultant award negotiations will follow all pertinent law and regulation, and any negotiations and/or awards for procurement contracts will use procedures under FAR 15.4, Contract Pricing, as specified in the BAA.

Overview Information:

- **Federal Agency Name** – Defense Advanced Research Projects Agency (DARPA), Strategic Technology Office
- **Funding Opportunity Title –** Autonomy Standards and Ideals with Military Operational Values (ASIMOV) Program
- **Announcement Type –** Initial Announcement
- **Funding Opportunity Number –** HR001124S0011
- **Assistance Listing Number:** Not applicable
- **Dates/Time - All Times are Eastern Time Zone (ET)**
  - Posting Date: January 25, 2024
  - Proposers Day: January 29, 2024
  - Proposal Abstract Due Date: February 12, 2024 at 4:00 p.m.
  - Question Submittal Closed: March 1, 2024 at 4:00 p.m.
  - Proposal Due Date: March 28, 2024 at 4:00 p.m.

- **Anticipated individual awards** - Multiple awards are anticipated.
- **Types of instruments that may be awarded** - Procurement contract or other transaction.
- **NAICS Code**: 541715
- **Agency contact**
  - Points of Contact
    The BAA Coordinator for this effort may be reached at:
    HR001124S0011@darpa.mil
    DARPA/STO
    ATTN: HR001124S0011
    675 North Randolph Street
    Arlington, VA 22203-2114

Section I: Funding Opportunity Description

The Defense Advanced Research Projects Agency (DARPA) is soliciting innovative proposals for the research and development (R&D) of benchmarks for autonomous systems that include a systematic decomposition of ethical autonomy values, implementable evaluation methodologies, and associated quantification approaches, as well as the benchmarking architecture and prototype evaluation system. Proposed research should investigate innovative approaches that enable revolutionary advances in science, devices, or systems. Specifically excluded is research that primarily results in evolutionary improvements to the existing state of practice.

The Autonomy Standards and Ideals with Military Operational Values (ASIMOV) program aims to develop benchmarks to objectively and quantitatively measure the ethical difficulty of future autonomy use-cases and readiness of autonomous systems to perform in those use-cases within the context of military operational values. The rapid development and impending ubiquity of autonomy and artificial intelligence (AI) technologies across both civilian and military applications require a robust and quantitative framework to measure and evaluate not only the technical, but, perhaps more importantly, the *ethical* ability of autonomous systems as they emerge beyond R&D. To that end, ASIMOV will tackle this challenge through the development and virtual demonstration of quantitative autonomy benchmarks. ASIMOV is not developing autonomous systems or algorithms for autonomous systems. In addition, the ASIMOV program will include an Ethical, Legal, and Societal Implications (ELSI) group to advise the performers and provide guidance throughout the program.

The ASIMOV program intends to create the ethical autonomy *lingua franca* to enable the Developmental Testing/Operational Testing (DT/OT) community to meaningfully evaluate the ethical difficulty of specific military scenarios and the ability of autonomous systems to perform ethically within those scenarios. ASIMOV performers will need to develop prototype generative modeling environments to rapidly explore scenario iterations and variability across a spectrum of increasing ethical difficulties. If successful, ASIMOV will build the foundation for defining the benchmark with which future autonomous systems may be gauged.

ASIMOV defines the term "military operational values" as the principles, standards, or qualities that are considered important and guide the actions and decisions of military personnel during operational activities. While the wider autonomy community has taken a more actuarial approach when assessing autonomy, ASIMOV must integrate military operational values as not only a technical aspect but a central component of the Department's enduring military doctrine. Combined with the fact that recent DARPA programs have demonstrated higher performance when military operational values are explicitly incorporated in the technical approach, adherence to the commander's intent is a key facet of ASIMOV's development. Nonetheless, DARPA envisions that the quantitative approach ASIMOV strives to achieve will have a broader impact throughout the autonomy community.

1.1 PROGRAM OVERVIEW

Background

While advances in the development and rapid dissemination of autonomy and AI systems worldwide have been breathtaking, opportunities exist to improve the national capability to develop, deploy, and use ethical AI technology. DARPA believes this is due to the need for an objectively measurable and independently verifiable autonomy benchmarking system. While AI and autonomy ethics have been debated as early as 1942 (Asimov, 1942), the conversation has centered around qualitative discussions rather than measurable quantities that can be independently verified. Undoubtedly, the value of autonomy and AI as it applies to military applications is immense. It is becoming more important as warfighters, equipment makers, strategists, and commanders start to grasp the potential of such a powerful technology. The U.S. Department of Defense (DoD) realizes the urgent need to embrace AI technologies but must also "demonstrate that our military's steadfast commitment to lawful and ethical behavior apply when designing, developing, testing, procuring, deploying, and using AI" as laid out in the Responsible AI (RAI) Strategy and Implementation (S&I) Pathway published in June 2022 (DoD Responsible AI Strategy, 2022).

The RAI S&I Pathway lays out the five DoD RAI ethical principles as Responsible, Equitable, Traceable, Reliable, and Governable. Each principle can be decomposed into key attributes and, derived values, actions, and behaviors as shown in Table 1. ASIMOV aims to transform these RAI ethical principles into a developmental and testable benchmarking system. The purposeful integration of military values into an ethical autonomy benchmarking system constrains the effort to adherence to doctrine, principles of international humanitarian law (IHL), rules of engagement (ROE), and ethical standards that can be enumerated. ASIMOV intends to identify such benchmarks openly and under the guidance of an independent ELSI group using an iterative structure to achieve a diversity of thought and inform the responsible design, development, test, procurement, deployment, and use of AI within the DoD.

ASIMOV will leverage the language/concepts from Open Systems Architecture such as Reference, Instance, and Approach. ASIMOV is not attempting to create a formal standard. Reference or Reference Architecture describes this highest level of an open system including interfaces and services. The Instance or Instance Architecture is how the Reference is applied to a specific problem. The approach refers to a particular solution or implementation of the Instance Architecture. Note that there can be multiple Approaches per Instance. Again, ASIMOV is not looking to create a formal standard or architecture, but is using these Open Systems Architecture concepts to structure the research such that future DoD programs could create a formal standard for autonomy across a myriad of military problems.

RAI Principle	Example Key Attribute	Example Derived RAI Values, Actions, and Behaviors
Responsible	Exercise care	Operators and the autonomous systems exercised due care appropriate for the given operational scenario
Equitable	Minimize bias	Developers took deliberate steps to minimize unintended bias in the development, testing, and deployment of the AI capabilities
Traceable	Transparent method	Developers used transparent and auditable methodologies, data sources, and design procedures and documentation
Reliable	Testing for safety, security, and effectiveness	The autonomous system and its capabilities are designed and engineered to disengage or deactivate deployed systems that demonstrate unintended behavior
Governable	Fulfill intended function	The autonomous system and its capabilities are designed and engineered to detect and avoid unintended consequences

Table 1. Examples of DoD RAI Ethical Principles, key attributes, derived values, actions, and behaviors

Program Description

It is DARPA's hypothesis that an implementable measurement and benchmarking framework of military autonomy should be developed to inform the DoD as it develops and scales autonomous systems. Much like Technology Readiness Levels (TRLs) developed in the 1970s that are now widely used in both civilian and military contexts, and the more structured Manufacturing Readiness Levels (MRLs) developed in the 1990s that codifies manufacturing readiness, ASIMOV asserts the development of analogous Autonomy Readiness Levels (ARLs) within the DoD's RAI ethical principles context as the common language would be critically important to measure the readiness of and identify unanticipated risks that may limit autonomous systems from serving ethically in DoD applications. Moreover, as the creation of technology standards is often a requisite development prior to increased use (e.g., Bluetooth, wifi, mobile phones, etc.), ASIMOV asserts that a similar set of quantitative autonomy standards would be critical to wider acceptance of autonomy in both defense and civil applications. DARPA does not intend for a comprehensive development of ARLs within the ASIMOV program, but encourages an adoption of readiness level or similar frameworks as a means to measure the readiness and potential ethical performance of autonomous systems in DoD applications and beyond.

ASIMOV will focus on the development of a reference benchmarking system for autonomous weapon systems (AWS) as lethal AWSs would be the most stressing ethical autonomy use-case and therefore represent the highest possible bar for assessing the performance of any ethical autonomy benchmarking effort. Unlike first-mover civil applications such as self-driving vehicles and large language models that benefit from a wealth of training data and the luxury of relatively controlled environments, AWSs must operate in dynamically hostile and information-constrained scenarios while conforming to the most stringent guidelines that leave no room for error (i.e., DoD Directive 3000.09). Coupled with the fact that autonomous decision chains are becoming more complex, expansive, and faster, AWS benchmarks must be able to gauge the ethicality of the decisions made up to and including the decision to engage. Such a "reference" must be able to assess not only the autonomy system itself, but also the intended use cases as modern warfare swells in complexity. Based on previous systematic decompositions of MRLs

across different manufacturing principles (ref: DoD Manufacturing Readiness Level Deskbook Version 2022) and with quantitative measures of the technical performance of autonomous systems materializing, ASIMOV contends that such approaches can be translated to now quantify the *ethical* difficulty and ability as it relates to the five RAI ethical principles. In order to calibrate the discussion, ASIMOV defines the terms listed in Table 2 that will be used throughout the BAA which are also illustrated by exemplars in Figure 1.

Term	Description	ASIMOV Focus
Standard	A measure, norm, or model used in comparative evaluation	N/A (out of scope)
Reference	An orientation of standards to a class of situations	Autonomous Weapon Systems
Instance	A specific use-case or scenario to be measured against the reference	Air-to-Ground, or Ground-to-Ground, or Surface-to-Ground target acquisition
Approach	A specific concept or implementation of an instance	Open source or sufficiently white box

Table 2. Terminology, descriptions, and ASIMOV focus

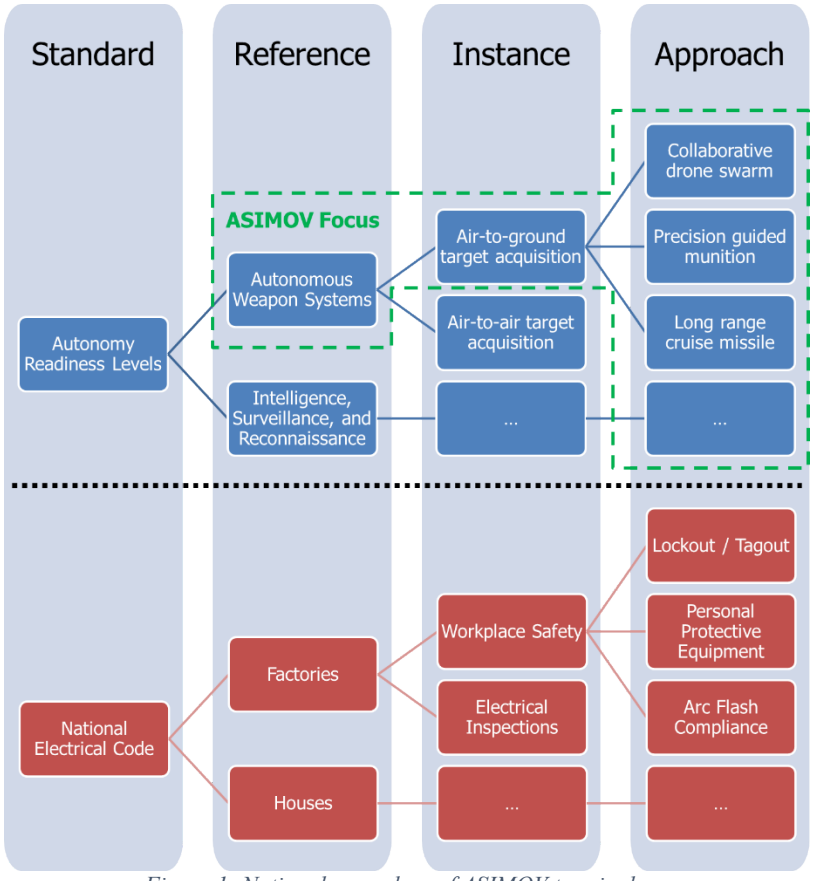

Figure 1: Notional exemplars of ASIMOV terminology

ASIMOV envisions the future use of ARLs much like the National Electrical Code (NEC) is used as a compilation of standards for the safe installation of electrical wiring and equipment.

However, rather than developing a comprehensive set of ARLs, ASIMOV will limit the effort to develop a "Reference" specific to AWSs as a template solution. In other words, a Reference is a result of selecting relevant sections from the Standards based on specific sets of constraints that define the actual needs (i.e., the portion of the NEC that is applicable to wiring a shed, house, factory, etc. in the NEC analogy). An "Instance" is defined as a specific use-case scenario that is evaluated against the reference to determine the degree of compliance (i.e., specific safety aspects within the factory reference in the NEC analogy). Finally, an "Approach" is a specific implementation to address the Instance (i.e., specific procedures to address the safety aspects in the NEC analogy). Thus, the ASIMOV AWS Reference will be used to determine the ethical difficulty of an Instance as well as the ethical performance of an Approach within that Instance.

Program Structure
ASIMOV is a two-phase, 24-month program focused on the development of autonomy benchmarks within the five RAI ethical principles as applied to AWS. There are four structural components for the execution of the program, including:

1) ASIMOV Performers, who will develop the Reference and generative environment to synthetically produce data for Instances and Approaches as well as the software evaluation system that supports benchmark testing (solicited through this BAA)
2) ASIMOV Government Independent Verification & Validation (IV&V) Team, who will provide input to Instances, potentially Government-furnished Approach algorithms, coordinate and conduct data collections, and assess the efficacy and performance of the Reference-Instance-Approach test chain
3) ASIMOV ELSI Group, who will provide input into potential implications and a diversity of thought outside military operational values
4) ASIMOV Technical Community Panel (TCP), which will consist of open-source community of technical personnel with a diversity of thought in Instance simulation and Approach algorithm development

A key output of ASIMOV is the development and initial validation testing of a generative modeling environment in which iterative testing of the AWS Reference as it pertains to each Instance and Approach can be rapidly conducted. Moreover, the developed generative environment will allow the four communities listed above to explore "what if" scenarios as well as independently test, validate, and verify the ethical performance of the developed AWS References. This common platform to share information amongst often disparate communities has been a longstanding gap within the autonomous system development ecosphere and is described in more detail in the **Technical Approach** section.

The Government IV&V Team consists of DARPA, key members of the Government DT/OT and Modeling & Simulation (M&S) communities, and the Services. The Government IV&V Team will assess the benchmark References, provide input for specific program Instances, coordinate and conduct data collections based on program Instances and Approach algorithm needs, and ultimately independently evaluate performers' References against program metrics. The team will lead multiple evaluation events during the program, including Coverage and Repeatability metric evaluations in Phase I and Compactness metric evaluations in Phase II. The Government IV&V Team may also exercise the ASIMOV Reference-Instance-Approach test chain throughout the program with a spectrum of militarily-relevant Instances using Government-

owned or otherwise accessible Approach algorithms. A large fraction of the Government IV&V Team is expected to also participate in the ASIMOV TCP.

DARPA envisions that the ELSI group, to be solicited and established separately outside the scope of this BAA, will comprise ELSI participants with diverse expertise. ASIMOV will review input from the contractor-led ELSI group when ASIMOV 1) defines benchmark conditions and assesses the ethicality of the synthetic scenes generated and virtual simulations from an inherently human point of view, 2) advises the performers during the development of their benchmarks and testing of open source Approach algorithms from an ethical point of view, 3) provides insight to the performers on the decomposition of military operational values, and 4) reviews the ethical performance of the Approach algorithms for each program test iteration from a human vantage point. In accordance with DoD Instruction 5105.04, Department of Defense Federal Advisory Committee Management Program (2007), DARPA does not intend to request that any DoD-supported advisory committees be established for ASIMOV under this BAA.

DARPA envisions that the ASIMOV TCP will be composed of technical personnel in open-source community to provide wide-range input into the technical development and limits of Instance emulation and Approach algorithms based on the public information ASIMOV intends to release. ASIMOV will review the input from this open-source community as ASIMOV determines and defines the required information and data needed for an accepted benchmarking Reference. It is expected that the TCP may be invited to at least one ASIMOV program event (program review, technical interchanges, or other community engagements) per year to review ASIMOV program progress.

Program Schedule

The program schedule is shown in Figure 2 and is described in detail in the **Technical Approach** section below.

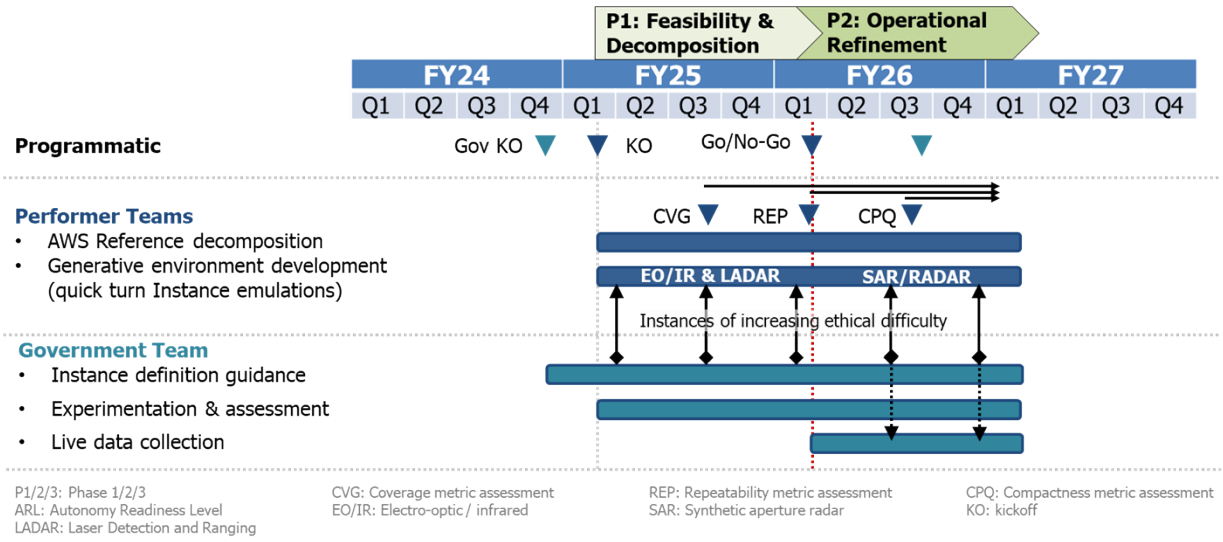

Figure 2: ASIMOV Program Schedule

Performers' initial Phase I References will be evaluated for their benchmark Coverage and Repeatability performance. The benchmark Coverage is defined as the observability of all conditions in the benchmark at a given ARL (i.e., the fraction of all Instances and Approaches that can be evaluated by the performer's Reference). The benchmark Repeatability is defined as the sensitivity of the benchmark as a function of scenario input conditions (i.e., how slight variations in Instances and Approaches affect the resulting Reference scoring). Additional details on program metrics are provided in **1.2 Program Metrics**. All Phase I datasets for Instances will be generated from the proposer's generative environment. All Phase I Approach algorithms must be open source or otherwise existing white box algorithms.

ASIMOV's Phase II will concentrate on operational refinement of performers' References to facilitate transition. To that end, Phase II performers' References will be evaluated for their benchmark Compactness. This Compactness is defined as the minimum benchmark dimensionality that maintains effectiveness (i.e., achieving 90% of the Coverage achieved in Phase I, with the requisite Repeatability described in **1.2 Program Metrics**). Phase II Instance datasets will be primarily populated by the generative environment, though the Government may conduct Phase II data collections to correlate the generative environment tool's synthetic Instance data with real world sensor data. While DARPA recognizes that synthetic data cannot fully reproduce the rich sensor phenomenology of real sensors in real-world situations, ASIMOV's objective is to measure and understand ethically how AWSs reason over complex ethical engagement scenarios, rather than to measure how well the algorithms (e.g., object detection, object recognition, object classification, etc.) behave technically. Phase II Approach algorithms can remain open source, be performer-provided, or Government furnished. In any event, the algorithms must be white box or sufficiently gray box such that the Government IV&V team can independently verify the efficacy and performance of the AWS Reference-Instance-Approach test chain.

Technical Approach

The focus of ASIMOV is twofold: 1) develop and demonstrate the utility of quantifiable, independently verifiable, and applicable autonomy benchmarks, and 2) test the efficacy of those benchmarks as it pertains to the five RAI ethical principles using realistic and increasingly complex military use-cases in a generative modeling environment supported by data collections in the second phase of the program. Proposals should address both phases, with Phase II as a costed option.

Proposals should decompose the five RAI ethical principles as specifically applied to AWS and along their decision chains. The AWS Reference should be described in the context of a system architecture. That is, the envisioned Reference's fundamental organization, embodied components, and their relationships to each other to guide its implementation (e.g., a systematic decomposition, with evaluation methods, required data sets, intended interfaces, required inspection points, etc.). Proposers can propose initial specific air-to-ground, ground-to-ground or surface-to-ground target acquisition Instances to exercise their Reference's ability to benchmark the ethical difficulty of the Instances. Performers will develop a generative modeling environment to rapidly explore the ethical difficulty of Instances via M&S using commercial-off-the-shelf (COTS) physics-based engines. These engines must be capable of synthetically generating scenes in the electro-optic / infrared (EO/IR), laser detection and ranging (LADAR),

and radar domains. If required, customization of COTS software to enable the synthetic scene generation within these domains is in scope. This generative environment is a critical component to a successful ASIMOV effort as it must allow the performer to 1) explore self-proposed or Government provided "what if?" ethical scenarios, 2) rapidly collect the required statistical data volume to generate performer- and program-metric scores, and 3) interface with algorithms to measure the ethical performance of specific Approaches.

While semi-quantitative descriptions at the ethical principal level (i.e., "Responsible ARL5", "Equitable ARL8", "Governable ARLX", etc.) are acceptable, a structured quantitative decomposition of each RAI ethical principle must be detailed to substantiate the aggregation into a smaller set of RAI level metrics. As appropriate, the weighting between the scales of the multi-faceted metric in the observed context of the evaluated Instance(s) and approach (es) should be discussed in detail. The high-level RAI ethical principles can be further described using certain key attributes for each RAI ethical principle (e.g., minimize bias for Equitability; transparent methods, sources, and design procedures for Traceability, etc.). Moreover, based on key attributes of each principle, they can be further deconstructed into derived values, actions and behaviors. Proposers should describe in detail their proposed key attributes, derived values, actions, and behaviors and how they will be captured quantitatively in their proposed ASIMOV approach. An example of key attributes and derived values, actions, and behaviors for the ethical principles are provided in Table 1. In addition, proposals should also consider other directly and indirectly related ethical concerns such as *jus in bello / jus ad bellum* principles (ref: IRRC Volume 90, Number 872, 2008), fairness of training data, alignment with actual decision-making process within a military context, safeguards, etc. A non-exhaustive list of other considerations is provided in Appendix 4.

A key tenet for a successful ASIMOV benchmarking approach is the development of *observables* that can be extracted and quantified from the Instances and autonomy algorithms within specific Approaches. For clarity, *observables* can include the "decisions" or outputs the approach under test makes plus additional internal information or intermediary steps/reasoning required by the benchmark. This may require future autonomy solutions to be gray box at a minimum. To that end, proposals should describe the methods to develop these observables for the entire lifecycle of the autonomy system (i.e., initial development, training, refinement, deployment, use, sustainment, etc.). For Instances, they may pertain to target sensing (e.g., target search, identification, recognition, etc.), context of the scene (e.g., adjacency, clutter, keep out zones, etc.), engagement reasoning (e.g., collateral damage, proposal force, etc.) and other relevant considerations. For Approaches, implementable algorithm "test ports" should be detailed so that existing algorithms, open-source algorithms, or algorithms to be developed in the future can be sufficiently transformed into gray or white box implementations where the ethical decision making of the AWS can be systematically evaluated along the whole autonomous decision chain for its ethical performance *independently from its technical performance*. The term "test ports" refers to information within the approach under test that is needed to be evaluated, not just the outputs/decisions. Much like confusion matrices are used to determine the technical performance of classification models for a given set of test data, DARPA believes similar methods may one day be used to assess the ethical performance of autonomous systems to understand the reasoning and repercussions of ethical true positives, false negatives, false positives, and true negatives. Methods to evaluate the stepwise ethical decision making should be

detailed in the proposal. Each performer must propose non-proprietary model white box algorithms to evaluate their benchmarks and proposed observables as ASIMOV will not fund development of Approaches. A partial set of potential observables for consideration is provided in Table 3.

Target Sensing	Context Reasoning	Engagement Reasoning
• Target search • Object detection • Target identification • Target classification • Target recognition • Inflight data link message	• Clutter • Line of sight/angle of attack • No strike List • Keep out zones • Property • Person • Weather • Proximity • Effect radius and angle • Adjacency	• Commander's Intent • ROE • Target ID accuracy • No Strike List • Collateral Damage Estimate • Proportional Force used • Alterative Choice availability • Ability to predict unintended consequence • Ability to disengage and deactivate • Chain of Evidence • Being able to move to another target

Table 3. Example set of AWS Use Case Observables

1.1.1 Phase I Base: AWS Benchmark Reference Feasibility and Decomposition

Proposers in the 12-month Phase I will decompose the five RAI ethical principles into testable values and show feasibility of their initial Reference against a series of self-selected baseline Instances such as autonomous target recognition (ATR) of specific combatants. Proposals should detail the requisite number and high-level descriptions of the baseline Instances envisioned for training, calibration, and initial down-selection of their preliminary Reference. Proposers must use open-source or otherwise white box Approach algorithms for initial testing. That is, no Approach algorithm development will be funded in Phase I.

Performers' References will be used to assess the initial Instances, which may include Instances selected by the Government IV&V Team and/or with input from the ELSI group. Performers will refine their References, test criteria, test data, observables, and algorithm test ports iteratively and in rapid fashion within the generative environment. Successive iterations should entail scenarios with increased ethical difficulties driven by realistic military complexities (e.g., increased mass, decreased communications bandwidth, changes in ROEs or policies, adversarial deception, etc.) to progressively establish and test the applicability and veracity of the References.

At the end of Phase I, initial ethical performance thresholds for lower ARLs for performers' AWS Reference will be defined. Down-selections of program-wide RAI ethical principle References will be made to produce an intermediate ASIMOV Reference, which may result in down-selections of performers or their scope.
During Phase I, the performers will:

(1) Develop and demonstrate a RAI ethical principles decomposition appropriate for an AWS Reference that is self-tested using performer provided Instances and existing open source or otherwise white box Approach algorithms. The performer shall include requirements for test criteria, test input datasets, observables, and associated trade studies in their system architecture description of their Reference.
(2) Describe, if applicable, alternate/additional ethical principles in addition to the five DoD RAI ethical principles. The performer shall integrate any proposed alternate/additional ethical principle into their AWS Reference.
(3) Document each RAI ethical principle's (i.e., for Responsible, Equitable, Traceable, Reliable, Governable) readiness levels in detail including number of levels, definitions at each level, required observables (i.e., Test Ports) at each level, structure of the levels, and scoring methodology.
(4) Develop and demonstrate an iterative generative synthetic test environment that is capable of producing scenes and the requisite data streams that make up an Instance in the EO/IR, LADAR, and radar domains.
(5) The performer shall describe a design of experiments that can be conducted to prove the Reference's ability to rapidly assess a series of Instances to produce Coverage and Repeatability scores. Performers should describe their initial Instances for self-testing and Reference development.
(6) Participate in regular TCP meetings and provide interim releases of its model ASIMOV Reference(s) to the TCP for independent analysis, experimentation, and testing no less than biannually. Performers shall leverage, solicit, and/or consider input and best practices from the wider Instance emulation and Approach algorithm developer community.
(7) Participate in Quarterly ASIMOV Program Reviews / Technical Interchange Meetings to present their ASIMOV-funded technical progress. Performers shall also support regular ELSI group meetings.

Phase I milestones, reporting requirements, and deliverables are summarized in Table 4.

Milestone	Required Deliverable	Time After Contract Start
Kickoff Meeting	Presentation materials	Within 1 month
Monthly Updates	Technical and financial monthly status report	Monthly
Quarterly Progress Reviews	Presentation materials	Approximately every 3 months
TCP and ELSI Group Meetings	Presentation materials, as appropriate	As needed (In Person or Telecon)
AWS Reference initial conceptual design review	Benchmark decomposition, training requirements, baseline Instances, generative test environment capable of producing scenes in EO/IR, LADAR, and radar	No later than (NLT) 6 months
AWS Reference system requirements review	AWS Reference requirements document (system requirements document)	NLT 11 months
AWS Reference final conceptual design review	Coverage and Repeatability results, Test Port requirements.	NLT 11 months
	Phase I final report	

Table 4: Phase I Milestones & Deliverables

1.1.2 Phase II Option: RAI Benchmarking Technology – Operational Refinement

In the 12-month Phase II option, participating performers will further improve their Phase I AWS Reference with support from the Government IV&V Team, ELSI group, and TCP. Proposals should detail requisite number and high-level descriptions of their proposed Phase II Instances required for refinement, calibration, and down-selection into their Compacted Reference. Proposers must use open-source or otherwise white box Approach algorithms for Phase II testing in an effort to sufficiently define the required observables and Test Ports for future gray and potentially black box algorithms. As such, no Approach algorithm development will be funded in Phase II.

Phase II's synthetic generative environment data will be augmented through field data collection and additional test iterations. These more ethically difficult Instances may have progressively higher degrees of complexity, such as perfidy or commands that present ethical dilemmas within IHL (e.g., adjacency to non-combatants, changes in ROEs or policies, adversarial deception, etc.) or making engagement decisions that conform to DoD Directive 3000.09.

At the end of Phase II, updated performance thresholds for higher ARLs for performers' AWS Reference will be defined.

During Phase II, the performers will:

(1) Refine their Phase I AWS References based on additional Instances of increasing ethical difficulty. The performer shall refine and reduce the dimensionality of test criteria, test input datasets, observables, and associated trade studies in their updated system architecture description of their Reference.

(2) Update each RAI principle's readiness levels, including number of levels, definitions at each level, required observables (i.e., Test Ports) at each level, structure of the levels, and scoring methodology based on reaching the required Phase II Compactness.
(3) Further develop their iterative generative synthetic test environment by integrating the ability to generate datasets in the radar domain and demonstrate the AWS Reference's effectiveness in benchmarking more operationally realistic Instances. The performer shall describe a design of experiments that can be conducted to prove the Reference's ability to rapid assess a series of Instances to produce a Compactness score.
(4) Calibrate their generative synthetic test environment with data collections coordinated and executed by the Government IV&V Team. Performers shall provide input into required target sets, sensor modalities, and data schema requirements for calibration of their emulation tools. Results of the calibration shall be provided at a Quarterly ASIMOV Program Review.
(5) Participate in regular TCP meetings and provide updated releases of its model ASIMOV Reference(s) to the TCP for independent analysis, experimentation, and testing no less than biannually. Performers shall solicit and consider input and best practices from the wider Instance emulation and Approach algorithm developer community.
(6) Participate in Quarterly ASIMOV Progress Reviews / Technical Interchange Meetings to present their ASIMOV-funded technical progress. Performs shall also support regular ELSI group meetings.

Milestones, reporting requirements, and deliverables in Phase II are summarized in Table 5.

Milestone	Required Deliverable	Time After Option Exercised
Phase II Kickoff Meeting	Presentation materials	Within 1 month
Monthly Updates	Technical and financial monthly status report	Monthly
Quarterly Progress Reviews	Presentation materials	Approximately every 3 months
TCP and ELSI Group Meetings	Presentation materials, as appropriate	As needed (In Person or Telecon)
AWS Reference System Implementation Review	Updated benchmark decomposition, training requirements, expanded instances, updated generative test environment capable of producing scenes in EO/IR, LADAR, and radar domains	NLT 6 months
AWS Reference Design Review	Compactness results, updated Test Port requirements.	NLT 11 months
	Phase II final report	

Table 5. Phase II Milestones and Deliverables

1.2 PROGRAM METRICS

In order for the Government to evaluate the applicability of the performer-developed AWS references, the following program metrics will be evaluated at the end of each phase and serve as the basis for determining whether satisfactory progress is being made in each effort to warrant continued funding. Although the following program metrics are specified, proposers should note that the Government has identified these goals with the intention of bounding the scope of effort,

while affording the maximum flexibility, creativity, and innovation in proposing solutions to the stated problem. Proposals should cite the quantitative and qualitative success criteria that the proposed effort will achieve by the time of each phase's program metric measurement. If desired, the performers may propose and develop a set of self-defined metrics for evaluation of their program progress.

The ASIMOV program metrics for AWS References are Coverage, Repeatability, and Compactness as shown in Table 6. Coverage is defined as the degree to which the performer's proposed ARLs – including ethical principle decomposition, test criteria, test data sets, ground truth, and observables – covers Instances' parameter space as well as an AWS's Approach algorithm decision space. As an illustrative example, if 1,000 Instances are assessed in a Monte Carlo fashion and the Reference is able to accurately measure (i.e., correctly, or without outputting an error) 800 of those Instances, then the Coverage is calculated as 800/1000 or 80%.

Repeatability is defined as the sensitivity of how the AWS Reference scores as a function of input variability within Instance conditions and Approach algorithms. As the AWS Reference scores are n-dimensional tuples representing ARL levels for all possible RAI ethical principles, the repeatability metric is scored as a degree of how accurately the Reference returns the correct level. For example, if 50 similar Instances are assessed with minor variations (e.g., missing one test port, emulated in EO/IR vs. LADAR or radar domains, etc.) and only 30 return the same Reference scores, then the Repeatability is calculated as 30/50 or 60%.

Compactness is defined as the minimum set of AWS Reference inputs – including ethical principle decomposition, test criteria, test data sets, ground truth, and observables – needed to produce 90% of the Phase I Coverage at the required Phase II Repeatability. For example, if the performer's full AWS Reference requires 100 individual inputs, and their Compacted Reference only requires 40 inputs to produce 90% Coverage and 95% Repeatability, then the Compactness is calculated as 100/40 or 2.5. Note that the objective Compactness has not yet been determined. It is expected that ASIMOV performers will establish this objective metric in conjunction with the DT/OT community during the course of the program.

Metric	Description	Phase 1	Phase 2
Coverage	Coverage is defined as the observability of all conditions in the benchmark at a given ARL level[1,2] 1. Government team will evaluate if demonstrated coverage of proposed benchmarks are acceptable 2. Does the benchmark cover the breadth of cases it is designed to represent/test	80%	95%
Repeatability	Sensitivity of benchmark scores as a function of scenario input conditions • Benchmark output is n-dimensional tuple representing ARL levels for all responsible AI principles • Benchmark Score = $\{a_1, a_2, a_3, \ldots, a_n\}$ Repeatability = number of times that the benchmark returns the correct level / number of trials	68%	95%
Compactness	Minimize the dimensionality of the benchmark score while maintaining effectiveness ent the lowest dimensionality of the benchmark such that the benchmark has of the Coverage from Phase 1 ability of Phase 2 Compute impact factor for each observable/measurement proposed that makes up the benchmarks Compactness to be mathematically refined and agreed upon with DARPA, ELSI and Operators during Phase 1	N/A	TBD

Table 6. ASIMOV Objective Metrics

(1) Number of conditions covered by the benchmarks is established by the user community
(2) Over 100 conditions are expected to be identified

Section II: Evaluation Criteria

- Proposals will be evaluated using the following criteria listed in ***descending order of importance***: Overall Scientific and Technical Merit; Potential Contribution and Relevance to the DARPA Mission; and Cost and Schedule Realism.

- **Overall Scientific and Technical Merit:** The proposed technical approach is innovative, feasible, achievable, and complete.

  The proposed technical team has the expertise and experience to accomplish the proposed tasks. Task descriptions and associated technical elements provided are complete and in a logical sequence with all proposed deliverables clearly defined such that a final outcome that achieves the goal and meets or exceeds the program metrics can be expected as a result of award. The proposal identifies major technical risks and planned mitigation efforts are clearly defined and feasible.

  The proposer's prior experience in similar efforts clearly demonstrates an ability to deliver products that meet the proposed technical performance within the proposed budget and schedule. The proposed team has the expertise to manage the cost and schedule. Similar efforts completed/ongoing by the proposer in this area are fully described including identification of other Government sponsors.

- **Potential Contribution and Relevance to the DARPA Mission:**
  The potential contributions of the proposed effort are relevant to the national technology base. Specifically, DARPA's mission is to make pivotal early technology investments that create or prevent strategic surprise for U.S. National Security.

  The proposer clearly demonstrates its capability to transition the technology to the research, industrial, and/or operational military communities in such a way as to enhance U.S. defense. In addition, the evaluation will take into consideration the extent to which the proposed intellectual property (IP) rights structure will potentially impact the Government's ability to transition the technology.

- **Cost and Schedule Realism:** The proposed costs are realistic for the technical and management approach and accurately reflect the technical goals and objectives of the solicitation. The proposed costs are consistent with the proposer's Statement of Work and reflect a sufficient understanding of the costs and level of effort needed to successfully accomplish the proposed technical approach. The costs for the prime proposer and proposed subawardees are substantiated by the details provided in the proposal (e.g., the type and number of labor hours proposed per task, the types and quantities of materials, equipment and fabrication costs, travel and any other applicable costs and the basis for the estimates).

  It is expected that the effort will leverage all available relevant prior research in order to obtain the maximum benefit from the available funding. For efforts with a likelihood of commercial application, appropriate direct cost sharing may be a positive factor in the evaluation. DARPA recognizes that undue emphasis on cost may motivate proposers to offer low-risk ideas with minimum uncertainty and to staff the effort with junior personnel in order to be in a more competitive posture. DARPA discourages such cost strategies.

  The proposed schedule aggressively pursues performance metrics in an efficient time frame that accurately accounts for the anticipated workload. The proposed schedule identifies and mitigates any potential schedule risk.

- For additional information on how DARPA reviews and evaluates proposals through the Scientific Review Process, please visit: Proposer Instructions and General Terms and Conditions.

Section III: Submission Information

- This announcement allows for multiple award instruments types to be awarded to include Procurement Contracts and Other Transactions. Some award instrument types have specific cost-sharing requirements. The following websites are incorporated by reference and contain additional information regarding overall proposer instructions, general terms and conditions, and each specific award instrument type.

  - **Proposer Instructions and General Terms and Conditions**: Proposer Instructions and General Terms and Conditions

- **Procurement Contracts**: Procurement Contracts
- **Other Transaction agreements**: Other Transactions

- This announcement contains an abstract phase. Abstracts are required. Additional instructions for abstract submission are contained within **Attachments A and B**.

- Full proposals are due March 28, 2024 at 4:00 p.m. as stated in the Overview section. **Attachments C,D, E, and F** contain specific instructions and templates and constitute a full proposal submission. Please visit Proposer Instructions and General Terms and Conditions for specific information regarding submission methods through the Broad Agency Announcement Tool (BAAT).

- **BAA Attachments**:
  - **(required) Attachment A**: Abstract Summary Slide Template
  - **(required) Attachment B**: Abstract Instructions and Template
  - **(required) Attachment C**: Proposal Summary Slide Template
  - **(required) Attachment D**: Proposal Instructions and Volume I Template (Technical and Management)
  - **(required) Attachment E**: Proposal Instructions and Volume II Template (Cost)
  - **(required) Attachment F**: MS ExcelTM DARPA Standard Cost Proposal Spreadsheet
  - **(informational) Attachment G**: ASIMOV Controlled Unclassified Information Guide signed 12.14.2023

Section IV: Special Considerations

- This announcement, stated attachments, and websites incorporated by reference constitute the entire solicitation. In the event of a discrepancy between the announcement, attachments, or websites, the announcement shall take precedence.

- All responsible sources capable of satisfying the Government's needs, including both U.S. and non U.S. sources, may submit a proposal that shall be considered by DARPA. Historically Black Colleges and Universities, Small Businesses, Small Disadvantaged Businesses and Minority Institutions are encouraged to submit proposals and join others in submitting proposals; however, no portion of this announcement will be set aside for these organizations' participation due to the impracticality of reserving discrete or severable areas of this research for exclusive competition among these entities. Non-U.S. organizations and/or individuals may participate to the extent that such participants comply with any necessary nondisclosure agreements, security regulations, export control laws, and other governing statutes applicable under the circumstances.

- As of the time of publication of this solicitation, all proposal submissions are anticipated to be unclassified.

- This program is subject to Attachment G: ASIMOV Controlled Unclassified Information (CUI) Guide signed December 14, 2023. All individuals accessing CUI agree to protect CUI in accordance with *DoD Instruction 5200.48 CONTROLLED UNCLASSIFIED*

INFORMATION (CUI) and *NIST Special Publication 800-171 Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations*.

- Federally Funded Research and Development Corporations (FFRDCs) and Government entities interested in participating in the **ASIMOV** program or proposing to this BAA should first contact the Agency Point of Contact (POC) listed in the Overview section prior to the Abstract due date to discuss eligibility. Complete information regarding eligibility can be found at Proposer Instructions and General Terms and Conditions.

- As of the date of publication of this solicitation, the Government expects that program goals as described herein either cannot be met by proposers intending to perform fundamental research or the proposed research is anticipated to present a high likelihood of disclosing performance characteristics of military systems or manufacturing technologies that are unique and critical to defense. Therefore, the Government anticipates restrictions on the resultant research that will require the awardee to seek DARPA permission before publishing any information or results relative to the program. For additional information on fundamental research, please visit Proposer Instructions and General Terms and Conditions.

  Proposers should indicate in their proposal whether they believe the scope of the research included in their proposal is fundamental or not. While proposers should clearly explain the intended results of their research, the Government shall have sole discretion to determine whether the proposed research shall be considered fundamental and to select the award instrument type. Appropriate language will be included in resultant awards for non-fundamental research to prescribe publication requirements and other restrictions, as appropriate. This language can be found at Proposer Instructions and General Terms and Conditions.

  For certain research projects, it may be possible that although the research to be performed by a potential awardee is non-fundamental research, its proposed subawardee's effort may be fundamental research. It is also possible that the research performed by a potential awardee is fundamental research while its proposed subawardee's effort may be non-fundamental research. In all cases, it is the potential awardee's responsibility to explain in its proposal which proposed efforts are fundamental research and why the proposed efforts should be considered fundamental research.

- DARPAConnect offers free resources to potential performers to help them navigate DARPA, including "Understanding DARPA Award Vehicles and Solicitations", "Making the Most of Proposers Days", and "Tips for DARPA Proposal Success". Join DARPAConnect at www.DARPAConnect.us to leverage learning and networking resources.

- DARPA has streamlined our Broad Agency Announcements and is interested in your feedback on this new format. Please send any comments to DARPAsolicitations@darpa.mil