



CDAO

Human Systems Integration Test and Evaluation of Artificial Intelligence-Enabled Capabilities

What to Consider in a Test & Evaluation Strategy

April 2024

Table of Contents

01

T&E Strategies for AIECs

pp. 2–6

Provides an overview of the framework for the T&E of DoD AIECs and the role of this document within CDAO's T&E of AIECs Guidance and Best Practices product series.

02

HSI T&E Is Important

pp. 7–15

Justifies the importance of HSI in a world of autonomy and AIECs and discusses how operationally relevant context impacts technology utilization.

03

Challenges of HSI T&E of AIECs

pp. 16–35

Introduces 13 HSI concepts and discusses their relevance to the T&E of DoD AIECs to provide non-HSI experts with a high-level understanding.

04

HSI T&E over the AIEC Lifecycle

pp. 36–47

Highlights where HSI T&E should be incorporated across the AIEC lifecycle, including during acquisition and sustainment.

05

Reflecting on HSI T&E of AIECs

pp. 48–50

Summarizes recommended improvements for the HSI T&E of AIECs, reflecting on the challenges posed by the inclusion of AI and how T&E changes vary over the AIEC lifecycle.

This document provides a foundational overview how leveraging artificial intelligence (AI) in DoD capabilities will influence human systems integration (HSI) test and evaluation (T&E) considerations in DoD T&E Strategies (TESS).

The HSI T&E of AI-enabled capabilities (AIECs) is necessary to build justified confidence that DoD warfighters can utilize their technology to execute their missions successfully.

T&E Strategies for AIECs

This Section:

- + Specifies the role of the current document within the larger framework
- + Provides an overview of the framework for the test and evaluation of AI-enabled capabilities produced by CDAO Assessment and Assurance

01



This document is part of a framework for the T&E of AI-enabled capabilities

CDAO Assessment and Assurance is creating a framework to provide guidance on how to test and evaluate (T&E) AI-enabled capabilities (AIECs).

What is the framework?

The T&E of AIEC Framework provides best practices and guidance on how to test and evaluate AIEC.

The framework is organized into four categories of testing and provides different types of resources to AIEC developers and working-level testers.

Why is it needed?

The DoD community for the T&E of AIEC comes from a variety of backgrounds.

The T&E of AIEC Framework promotes a shared understanding between AIEC experts new to T&E and to T&E experts new to AIEC.

What is this document?

This document discusses what aspects of human systems integration (HSI) T&E to consider in a Test and Evaluation Strategy (TES) for an AIEC.

It is intended to help AIEC developers and working-level testers understand the importance of HSI in a world of autonomy and AIECs.

This document provides:

- ✓ **Guidance and best practices**
- ✓ **A primer on HSI T&E of AIECs**
- ✓ **Strategy-level T&E considerations**
- ✓ **HSI-specific T&E topics**

This document does NOT provide:

- ✗ **Binding policy and requirements**
- ✗ **A comprehensive HSI T&E guide**
- ✗ **Detailed T&E implementation**
- ✗ **T&E at the algorithm level**



CDAO's T&E of AIEC framework is organized into four focus areas

While these T&E focus areas help break critical aspects of T&E into digestible pieces, they are neither mutually exclusive nor cleanly delineated in real testing.



Operational T&E (OT&E)

Evaluating an AIEC performing representative missions within an operationally realistic environment against a realistic adversary.



Human Systems Integration (HSI) T&E

Evaluating an AIEC's ability to help stakeholders observe and orient to their environment, make informed decisions, and carry out their missions.



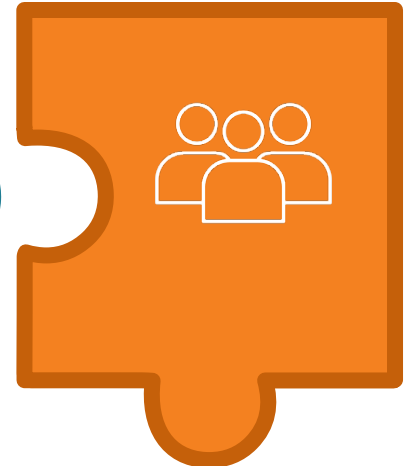
Systems Integration (SI) T&E

Evaluating an AI component within its larger system to ensure that the AIEC functions as a holistic unit and identify its limitations and risks.



AI Model T&E

Evaluating and documenting AI models and data across performance dimensions informed by system and mission constraints.



This document covers the HSI T&E focus area



CDAO is developing a series of products that address critical T&E needs

Part 1 is designed to help testers understand core T&E concepts so that working-level testers can write and assess test and evaluation strategies for AI-enabled capabilities

This document focuses on Part 1



1 | Write and assess T&E Strategies

Provides a high-level overview of critical T&E concepts that will be influenced by the inclusion of AI models in the system under test.

Supports testers and developers as they write TESs and assess whether the TES is committed to the right evaluations.



2 | Write and assess Detailed Test Plans

Provides guidance for implementation of T&E concepts introduced in Part 1; highlights promising paths forward for unsolved challenges.

Supports testers and developers as they develop and implement detailed test plans that capture mission objectives.



3 | Engage with other DoD T&E stakeholders

Provides frameworks outlining how T&E is critical to fielding trustworthy AIECs across DoD acquisition pathways and mission applications.

Supports testers and developers as they advocate for policy and investments that address DoD T&E shortcomings.



4 | Execute tests and rigorously analyze results

Provides resources such as templates, validated measurement instruments, and automated analysis tools.

Supports testers and developers by streamlining and automating common T&E activities with tailorable tools.



What is a Test & Evaluation Strategy?

A high-level document in DoD acquisitions that guides test planning and execution.



Captures the mission(s) a capability is intended to perform and all hardware and interfacing systems in the test design.



Identifies and prioritizes assessment areas to inform test team data requirements to support major program decisions.



Specifies the resources required to conduct T&E and shortfalls in resourcing that will require investments.



Describes the test events and activities necessary to evaluate the system and support acquisition, technical, and program decisions.



Learn More

You can read more about DoD TESs at
<https://www.test-evaluation.osd.mil/T-E-Enterprise-Guidebook/>



HSI T&E is important

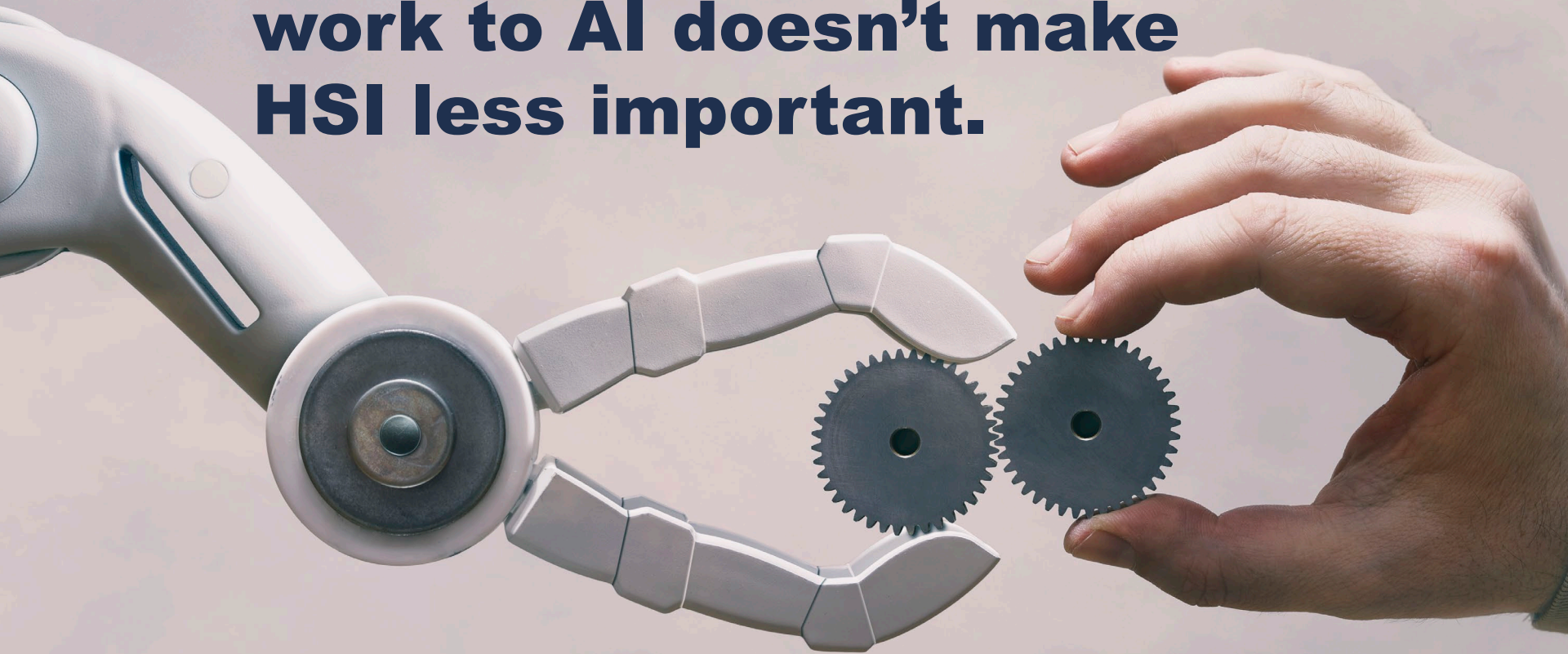
This Section:

- + Provides a brief overview of human systems integration (HSI) and explains why it is critical to do it well
- + Justifies the importance of HSI in a world of autonomy and AIECs and discusses how operationally relevant context impacts technology utilization
- + Introduces the OODA loop as a heuristic for thinking through program-specific HSI implementation

02



**Offloading warfighter
work to AI doesn't make
HSI less important.**



**It makes it more
important than ever.**



Human systems integration is important

Human systems integration (HSI): Design, development, and sustainment practices that ensure that warfighters can efficiently, effectively, and safely leverage technologies to accomplish tasks.

See [DoD 5000.02 E7](#) (page 79) for formal requirements.

Mission success is not achieved by systems operating in isolation.

Every program must characterize the HSI of their system in context with the user in context. HSI practitioners work toward making sure that technology augments humans' strengths and mitigates their weaknesses in the service of achieving mission objectives. HSI lessons have often been learned through catastrophe and best practices are written in blood. Ignoring these lessons and best practices invites huge risk in the world of military AI. Much of what has been learned comes from complex automation (e.g., commercial aviation) and can be repurposed for AIECs. DoD should do so wherever possible.

Warfighters are not the only users of these novel technologies.

Many of the HSI concepts presented in this document will be critical for many system stakeholders as well. Developers, testers, acquisition leads, executive-level decision-makers, and field commanders will all need varying levels of system understanding—i.e., mental models—to do their respective jobs. Onboard instrumentation might provide information to help warfighters maintain situational awareness, but other stakeholders will have different needs from that instrumentation. The details will vary, but testers should keep in mind that many stakeholders, including themselves, must be able to effectively interact with the system.

What is “human-machine teaming (HMT)” and how does it fit into this framework?

HMT is a special case of HSI and for T&E purposes we define it as the human and machine (1) pursuing the same goal, (2) affecting the current state, and (3) coordinating actions. While many of the HSI concepts discussed in this framework are relevant for understanding and evaluating HMT, they are not sufficient. True teamwork will involve new challenges that are not within the scope of this document.

If the interactions between the system and warfighter meet the three requirements above, testers can refer to part one of [IDA's HMT framework](#) for more information about relevant metrics for the T&E of HMT.



HSI evaluations must consider more than interfaces and ergonomics

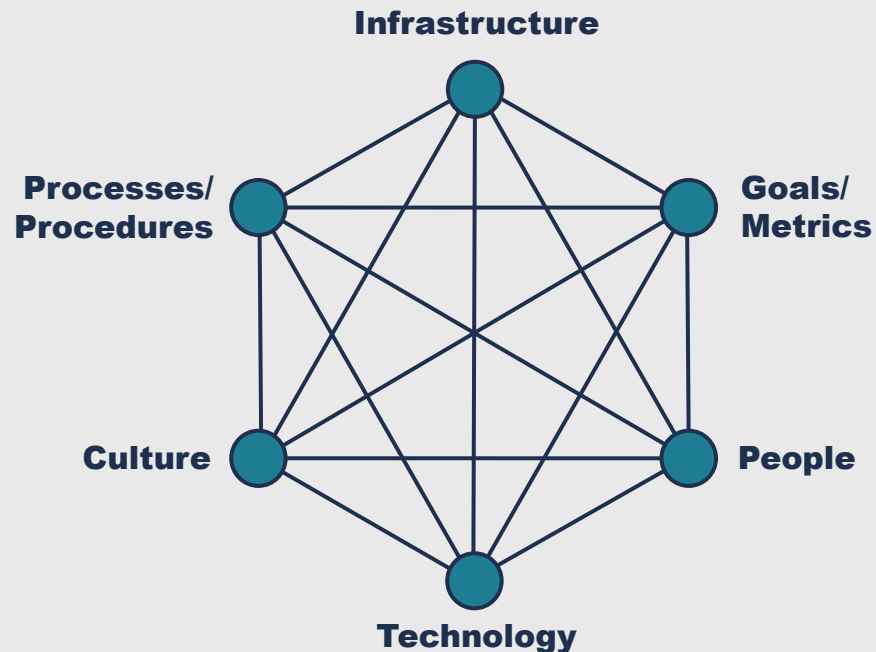
A holistic understanding of warfighters and systems in context will enable testers to plan and execute operationally representative tests and pick relevant HSI metrics.

HSI is often discussed narrowly, and many outside the field often focus on ergonomics and interfaces.

Interface design is still critical to creating usable technology; design choices ranging from information presentation to button size can all impact the usability of the technology.

Yet while both ergonomics and interface design are important, focusing only on the surface-level components of technology will be insufficient to achieve high-quality HSI.

Warfighters exist and work within socio-technical systems, and HSI designs and evaluations should consider mission performance in this context. A clever interface on top of a poorly designed system will not render the system usable.



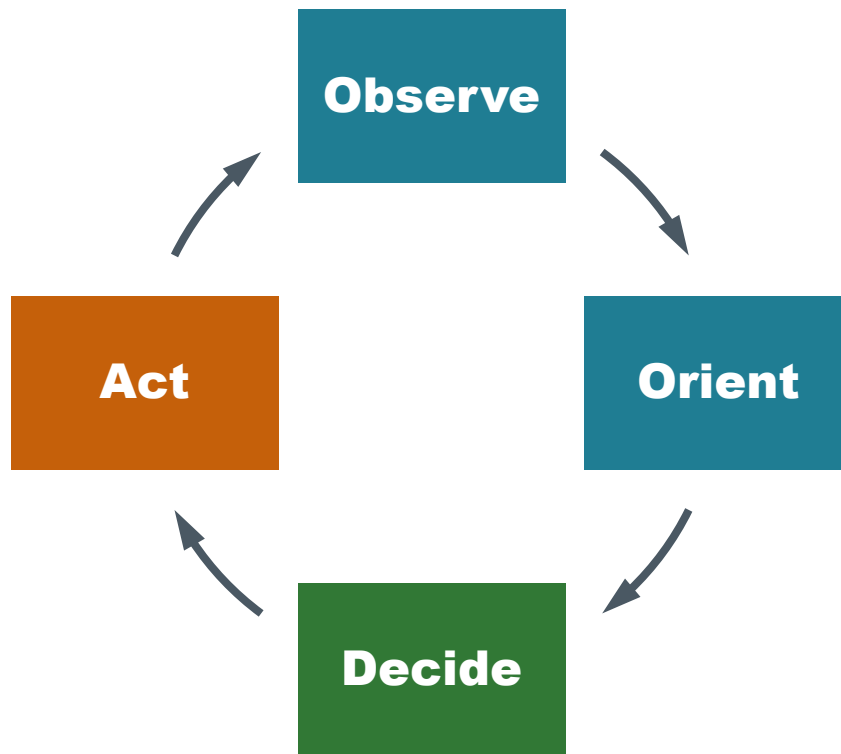
Socio-technical System

An interactive system formed by humans working with technology within an environmental context that includes goals and incentive structures, procedures, infrastructure, and workplace culture.



Testers need to think through how AIECs will augment or transform warfighters' work

It is helpful to breakdown our warfighters' work into a framework that allows us to conceptual tie DoD missions to HSI concepts in language familiar to DoD testers.



To decide what HSI concerns need to be emphasized, testers must understand how technology alters changes their work. For many in DoD, Boyd's Observe-Orient-Decide-Act (OODA) loop is a familiar framing.

When warfighters are assigned a mission, they are given a set of explicit (and implicit) goals about how they should try to change the environment. They have some idea about what needs to be achieved in order to say "mission accomplished."

To reach this end state, warfighters must:

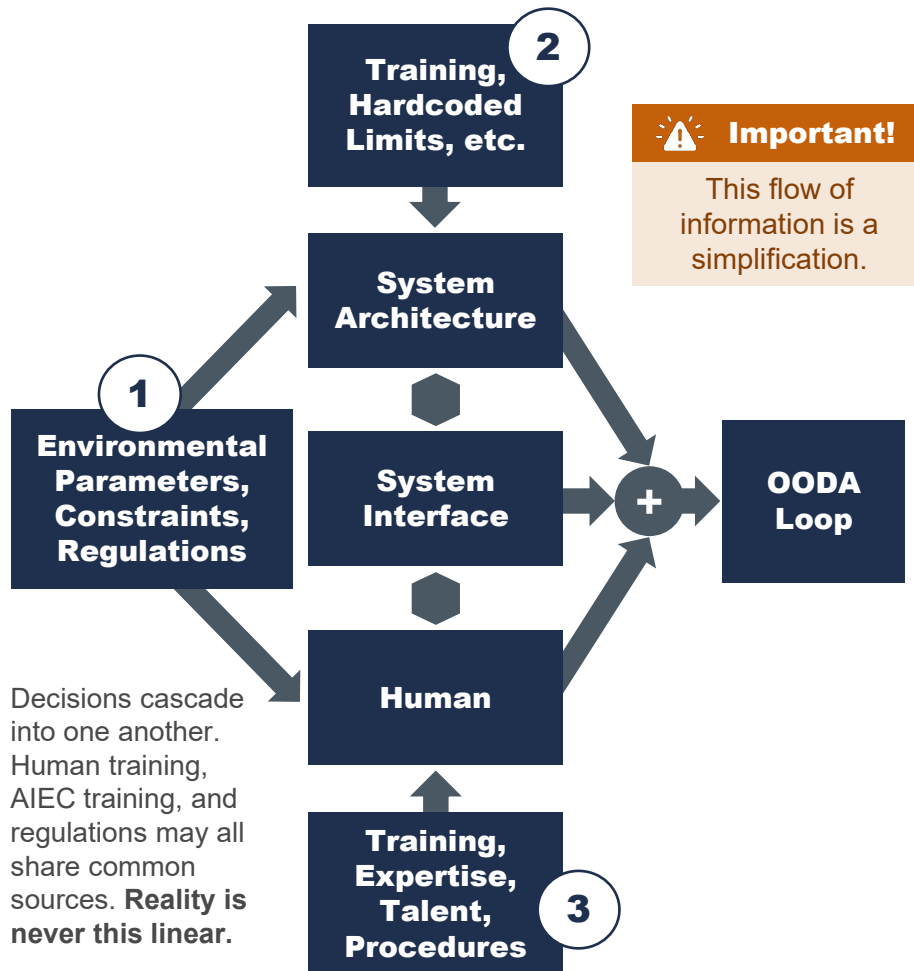
1. **Observe** their current environment.
2. Identify and interpret (**orient**) the goal-relevant information for the current state of the environment.
3. Leverage this information to **decide** on a course of action that will bring them closer to their goal.
4. Execute their intended **action**.

Because this last step will change the environment, operators loop through these steps as they try to accomplish their objectives.



The warfighter and automation do not exist in a vacuum; they must be evaluated in context

HSI goes beyond the human and the system. Testers should consider what affects the system, what affects the human, and how external factors influence each other.



1

Mission performance depends on external environmental factors.

At various points, the system and the warfighter must be able to respond to how the environment enables or constrains their shared performance through diverse factors, from regulations to weather conditions to data quality. **Testers should consider external context when designing HSI evaluations.**

2

System architecture will impact the warfighter's ability to successfully employ it.

A complex, unpredictable system cannot be transformed into an understandable system solely through interface design. The architecture will impact the warfighter's ability to understand and predict system behavior. **Testers should consider warfighters' mental models of the system during T&E.**

3

Warfighters are informed by their training and procedures.

The warfighter's interactions with the system are influenced by external forces. Their training, experience, and inherent talent, along with procedural support, will impact their ability to achieve mission success. **HSI evaluations should use representative populations—not “golden crews.”**



The OODA loop is like a generic user story

OODA is a useful framework for HSI T&E because most HSI concepts directly relate to performance of one or more of the decision-making stages of the OODA loop.

To develop an adequate evaluation plan, testers must understand what warfighters need from their technologies. OODA provides a familiar way to think through these issues. Testers should consider what stages their system augments, what downstream consequences this augmentation could have on the mission, and the resulting warfighter needs that must be tested. This will help in choosing test scenarios and implementing metrics.

DoD is comfortable developing and testing technologies that assist warfighters during the **Observe** and **Act** stages (e.g., radar and fly-by-wire). AI promises to massively expand augmentation at **Orient** and **Decide**. AIECs will perform all of these steps, and teaming systems will have complex dynamic interactions across these loops. Testers may be able to borrow approaches from systems that affect similar task components.

Below maps warfighter needs to the decision stages of the OODA loop

Observe & Orient

“I have to understand and predict the situation. Tell me what I need to know, when I need to know it, in a way that I understand.”

Decide

“I need to be able to make good decisions about where and how to use this system.”

Act

“I have to execute my decision. Make it easy to get the system to do what I intend it to do.”

This is not a strict mapping, but it can help testers think through HSI T&E issues



The warfighter and HSI remain vital in a world of autonomy and AIECs

Many assume that as AIECs automate more and more, humans will lose relevance.

This assumption is flawed.

Human interaction with technology can be complicated or undermined in a variety of ways. The complex nature of a task, the brittleness of a system, or a poorly designed user interface all introduce challenges for our warfighters as they try to leverage their technology to achieve mission objectives.

More complex AIECs will demand more complex interactions from warfighters, making HSI errors more likely. Along another dimension, increasing system autonomy will reduce warfighter touchpoints, leaving fewer opportunities to troubleshoot or correct erroneous inputs. Together, implementing good HSI design and appropriately evaluating those choices are more important than ever.

Most AIECs fielded in the near future will require engagement from warfighters.

As with traditional systems, it will be critical for AIECs to make information easily accessible, relevant, and understandable to the warfighter. However, when some envision AIECs they imagine a system in which warfighters will be able to “set it and forget it.” But even these advanced systems—which are not representative of near-term capabilities—will require a warfighter to initialize the task.

Consider an autonomous underwater minesweeper. With a more traditional system, the warfighter has opportunities to correct flawed inputs and adapt to unforeseen events. Conversely, a warfighter tasked with initializing the autonomous system needs a deeper understanding of the system, the environment, and the mission to provide informed inputs to achieve their goals.

Novel AIECs combined with evolving HSI will create challenges for HSI T&E.

Increasing system complexity and autonomy changes how warfighters interact with systems. It will be critical to evaluate warfighters’ ability to understand and predict system behavior. Testers will have to consider how to interpret mission performance evaluations, where system behaviors are tied to both warfighter inputs and the AI’s own internal decision-making.

Section 04 provides insight into how AIECs introduce new challenges for HSI T&E and highlights how failing to account for this evolving need can result in mission performance degradation or outright failure. A deeper discussion of the new challenges posed by AIECs to the T&E community—both generally and for HSI—can be found in IDA’s report: “Trustworthy Autonomy: A Roadmap to Assurance.”



This framework supports not only AIEC T&E but also HSI T&E of software-intensive systems

Many HSI challenges with AIEC already exist to varying degrees with traditional systems; programs without any AI components can also leverage this framework.

While AIECs introduce novel HSI T&E challenges, some solutions are found in the past.

The surging interest in AIECs has cast light on the many challenges—HSI and otherwise—that DoD faces in realizing these technical capabilities. But many of these challenges are not necessarily novel to AIECs and are currently faced by many software-intensive programs.

Automation has been redefining the human contribution to mission success for decades. This evolution of human work has led to best practices for considering HSI throughout the design process, allocating work between humans and automated agents, and—of course—leveraging T&E to provide an assurance case that our warfighters can work with their technology to achieve their mission goals.

Many HSI concepts in this framework are already required by DOT&E for TESSs.

Testers may already be familiar with some of the concepts discussed in this framework. Workload, trust, usability, and training quality are explicitly named in DOT&E guidance for qualitative and quantitative HSI evaluations.

This framework focuses on concepts most relevant to AI and how new technologies require updates to our evaluation methods. Most of the content, however, is relevant to T&E of current software-intensive systems. Specifically, [Section 04](#) provides “one-pagers” that summarize 13 HSI concepts. Testers can leverage these summaries as quick refreshers on HSI concepts and measurements or as a gateway to the research and tools that exist for that topic.

There is growing interest in HSI, but the T&E community often falls short.

Despite the recent emphasis on HSI and its critical contribution to the successful fielding of technology, it is often an afterthought in TESSs. It is not enough to just list “surveys” and “interviews” in TESSs as an HSI box-checking exercise. Even for non-AI systems, testers can and should leverage [Section 04](#) and [Section 05](#) to make sure that their TESSs explicitly name which HSI concepts should be included and how the evaluation is being triangulated.

The good news is that while some in the DoD test community have been historically resistant to improving HSI evaluation practices, when forced to change for the exacerbated AI challenges, it likely will become easier to implement these practices for software-intensive systems as well.



Challenges of HSI T&E of AIECs

This Section:

- + Introduces 13 distinct HSI concepts that are relevant to the T&E of AIECs
- + Presents each HSI concept on a single “one-pager” intended to provide non-HSI experts a high-level understanding
- + Maps each HSI concept to its related warfighter need and OODA loop stage (discussed in **Section 02**)

03

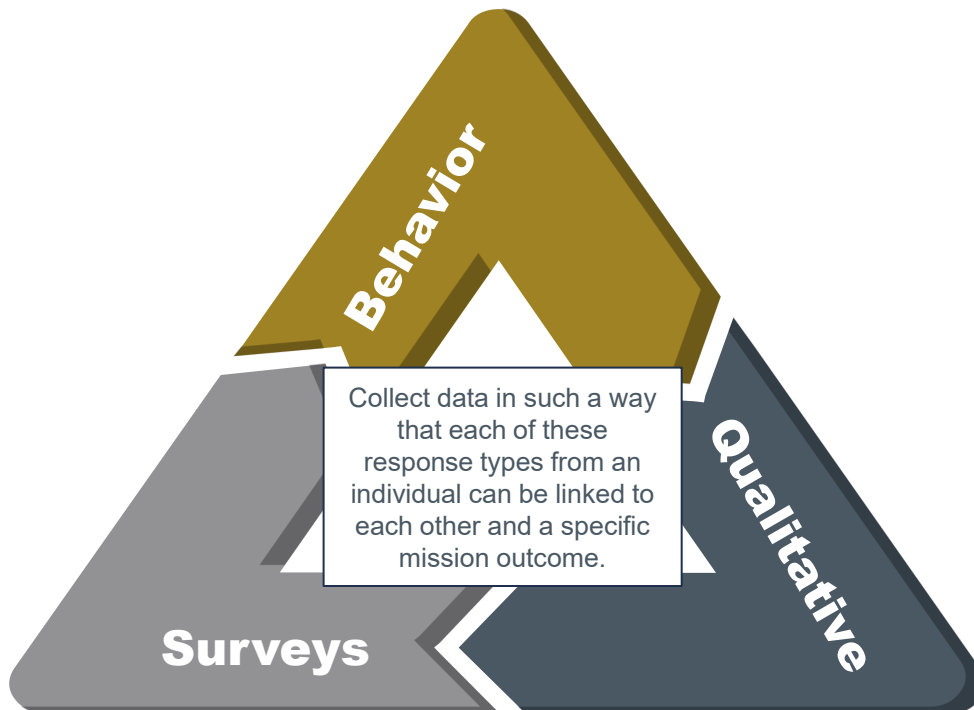


HSI T&E involves more than just surveys

Testers should leverage behavioral, survey, and qualitative methods. Each method has strengths and weaknesses, and it is best to use all three together in “triangulation.”

Triangulation

Statistically link HSI concepts to **objective, replicable operational effects** while placing these trends in the **human-interpretable context of a mission narrative**.



1

Behaviors are quantifiable human responses, including physiological ones.

PROS: Objective, accurate, and relevant; instrumentation allows automation.

CONS: HSI concept causation is nearly impossible to establish from behavior alone; see “reverse inference problem.”

2

Surveys quantify subjective experiences.

PROS: Quantifies a specific HSI concept so it can be statistically modeled.

CONS: Interpretation of relevance is difficult from survey alone. Getting accurate measurement requires a lot of work. Not very detailed.

3

Qualitative methods (i.e., interviews, focus groups, and comments) provide context.

PROS: Flexible and open-ended; details, context, and unexpected information can be understood.

CONS: Establishing broader trends from qualitative information alone is difficult. Labor-intensive and requires expertise.



How to use the HSI Concept “One Pagers”

Each HSI concept is presented in a “one-pager.”

This framework identifies 13 different HSI concepts that impact a warfighter’s ability to meaningfully leverage their AIEC, and should accordingly be included in a TES.

Use this section to write or review a TES so that it includes core HSI concepts relevant to the T&E of AIECs.

How should I use this section?

Identify core concepts: We identify the critical HSI concepts to consider when testing and evaluating AIECs.

Find “Google-able” terms: Each concept one-pager includes the formal name and definition. Beyond being informative, this formal language provides the keywords needed to search for additional resources.

Learn to interpret informal language: Because most TESs will not have input from HSI experts, one-pagers provide overviews and AI-specific concerns for testers to assess whether a TES has included relevant HSI concepts with different, informal language.

Understand the need to test: We explain how each HSI concept can either empower or undermine the effective, safe, or ethical employment of these novel systems.

What are the limitations of this section?

It is not an exhaustive product: While the core HSI concepts included in this product highlight key issues that testers should focus on, please be aware that this list is not complete. While more nuanced concepts and implementation guidance will be discussed in future “guidebook” and “deep dives”, no product in this series will exhaustively list all HSI concerns. Additionally, the concepts summary are limited to a single page, but in reality, most of these concepts span entire research communities.

Not all TESs will include all outlined HSI concepts: Every TES will not include all 13 HSI concepts in this framework right away. Some will have to prioritize resources, and some concepts may be less relevant for some systems.



HSI Concept One-Pagers

This subsection introduces 7 HSI concepts using the one-pager layout. These concepts have been grouped together because they are most related to a warfighter's need to perceive, understand, and anticipate environmental and system context.

**Warfighter
Need**

**"I have to understand and predict the situation.
Tell me what I need to know, when I need to
know it, in a way that I understand."**

**OODA Loop
Decision Stages**

Observe & Orient

**DoD Ethical
Principle**

Traceability

HSI Concepts

1. **Mental Models & Predicting System Behavior**
2. **Boundary Awareness**
3. **Information Quality: Objectivity**
4. **Information Quality: Utility**
5. **Information Quality: Interpretability**
6. **Situational Awareness**
7. **Explainable AI (XAI)**





Mental Models

Humans form mental models of processes in the world, and when applied to automation these models allow them to infer the current state of a system from incomplete information and make predictions about future states.

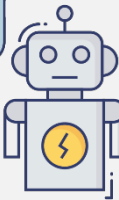
How is it relevant to testing?

Mental models (MMs) for AIEC will require an understanding of the system's decision-making processes—how operational conditions causally affect system behavior. MMs need to usefully predict behavior but not necessarily be perfectly “correct,” and different stakeholders will need different MMs. MMs are developed as the operator gains knowledge of and expertise with the system. Operators with well-developed MMs process information more efficiently and are better at leveraging relevant information while ignoring irrelevant information. While this filtering enables high performance, experts' MMs can become rigid, resulting in cognitive inflexibility and confirmation bias.

TESs should commit to evaluating the MMs operators develop and their ability to predict system behavior.



The operator's MM of a vehicle classification AIEC is that it “sees” enemy personnel carriers.



In reality, the AIEC sees only features, such as vehicle size, wheel count, and protrusions.



The operator would not predict that the AIEC would “see” this truck as a threat.

What could go wrong?

An FMV classifier is programmed to always return its “best guess,” even if all categories are low probability, but the operator believes the system can say “I don't know.” This poor MM results in the operator believing the system is confident in its “unarmored personnel carrier” classification and authorizing fires on a school bus.

How can AI make it harder?

Although the goal of building a mental model is not to have the perfect one—just a model that lets you usefully predict system behavior—the complexity and brittleness of modern AI will make it difficult to create human-understandable MMs that still sufficiently protect against catastrophic edge cases. There are many poorly understood nuances of AI algorithms that will cause operators' MMs to insufficiently predict system behavior. Operator MMs update over time, and MMs seen in test users may not be representative of those in later fielded users with more experience. Additionally, many AIs will be updated frequently (possibly even in real time), meaning that operators will be handed a (perhaps impossible) task to update their MMs frequently and accurately.

What are the state-of-the-art measurements?

Behavioral

Knowledge tests; risk assessments of operator's behaviors; accuracy of warfighter predictions of AI; controlled experiments to see how information changes predictions

Surveys

Some MM measures exist, but they are not well suited to military T&E. New scales should be developed and validated for this purpose.

Qualitative

User interviews



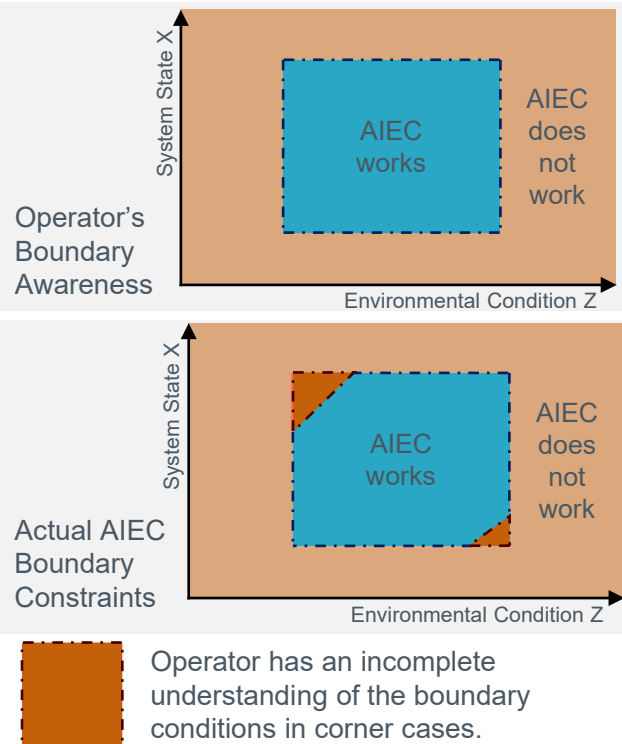
Boundary Awareness

This is the operator's understanding of the performance envelope of his or her system—i.e., its capabilities, limitations, and boundary conditions.

How is it relevant to testing?

Operators need to know the capabilities and limits of their system—that is, its operational boundaries. For example, pilots need to understand where an aircraft will stall in order to employ it effectively and safely. This awareness is related to their mental model of the system. Boundary awareness is related to but distinct from situational awareness; the former is awareness of system properties, and the latter is awareness of the current state (ideally relative to those system properties). Operators need to know where the system will consistently perform well or poorly, and where there is uncertainty, in order to properly calibrate their trust.

TESSs should commit to evaluating operators' boundary awareness and whether observed understanding is likely for typical operators.



What could go wrong?

A helicopter pilot is unaware that their threat recognition system does not work well in a forested environment, and their over-trust and failure to increase their vigilance leads to them being shot down.

How can AI make it harder?

To assess operators' boundary awareness, testers must know what those boundaries are. This will be harder to know than in standard systems. Unlike typical physics-bounded performance envelopes, decision processes usually have higher numbers of causally important factors. Those factors may be different among systems even for the same task. They might not be known at the start; the relationships to performance can be complex, non-linear, or discontinuous; and they might be hidden behind black boxes or proprietary screens. Understanding these systems will require experimentation that is not currently standard in acquisition pathways.

What are the state-of-the-art measurements?

Behavioral

Knowledge tests; risk assessment of operator's behaviors

Surveys

[Perceived risk scales](#)

Qualitative

User interviews



Information Quality: Objectivity

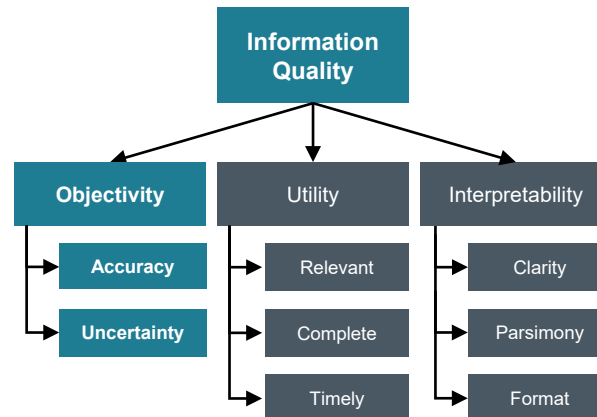
Information quality is determined by *objectivity*, *utility*, and *interpretability*. Objectivity focuses on whether the information sufficiently reflects reality for the user's needs, and whether the uncertainty of the reporter's estimate is communicated.

How is it relevant to testing?

To make appropriate decisions, operators must have good information about both their environment and systems. If this information is communicated rather than directly obtained, evaluating quality becomes critical. A key principle is that the quality dimensions should be assessed relative to task requirements, not at an absolute level. Warfighters never operate on perfect information, so testers must assess whether information is accurate enough for the task at hand. Situational factors change what is good enough (e.g., speed vs. accuracy), so a single accuracy requirement and evaluation will be insufficient.

TESs should commit to comparing the information accuracy and uncertainty provided versus warfighter needs across operational conditions.

Objectivity can be further broken down into **accuracy & uncertainty**.



Note: When describing human sources of information objectivity, there are additional criteria not included in this framework.

What could go wrong?

An operator does not understand that their target recognition system had a large amount of uncertainty when it identified a school bus as an enemy troop transport.

How can AI make it harder?

Testers must establish that the system is communicating information that correctly reflects the current state of the environment and/or system. However, AI also introduces the need to assess, at a meta-level, whether the operator (and sometimes the system itself) understands the uncertainty associated with the information. When operators collect information themselves, they can make this judgment on their own. AI capabilities can often remove operators from the context needed to make that assessment. AIECs must provide relevant information for users to judge that level of uncertainty.

What are the state-of-the-art measurements?

Behavioral

Automation to Interface instrumentation; working memory probes; user errors; task completion

Surveys

Some InfoQ measures exist, but they are not well suited to combat systems.

New scales should be developed and validated for this purpose.

Qualitative

User interviews



Information Quality: Utility

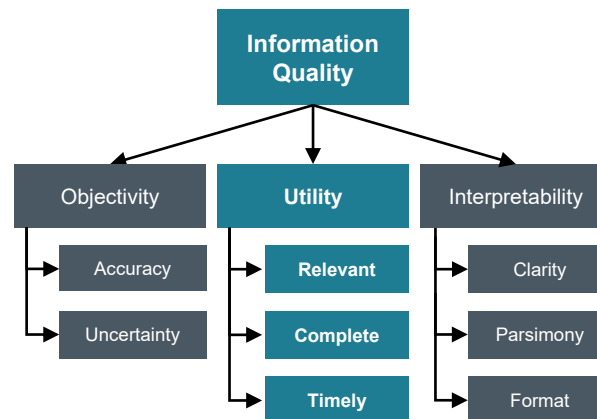
Information quality is determined by *objectivity*, *utility*, and *interpretability*. Utility focuses on how useful the information provided is for successfully completing the current task.

How is it relevant to testing?

The critical elements of information utility for AI are relevance, completeness, and timeliness. To be useful, the information communicated must be (1) relevant to the current task, (2) complete with all of the necessary bits of information included, and (3) delivered in a timely enough span to be actionable. All of these can be difficult to design and test, as the adequacy of these elements changes with the situation and task. What is adequate under one set of circumstances is different elsewhere. Completeness needs to be calibrated, as it is easy to overwhelm someone with too much information (some frameworks put parsimony as an aspect of completeness).

TESs should commit to testing information utility with representative warfighters in both DT and OT.

Utility can be further broken down into information relevance, completeness, and timeliness.



What could go wrong?

By the time a machine learning fault-recognition system onboard an aircraft has enough data to identify the cause of a flameout, it is too late to take corrective action.

How can AI make it harder?

Some information is more easily processed with modern machine learning techniques than others. Technological hurdles discovered mid-development may lead to design redirects, down scoping, or function reallocation. There are many ways in which the information or metrics that developers have chosen to communicate may be incomplete, irrelevant, or out of sync with the task. Testers need to assess whether users find operational value in the information conveyed. Additionally, Explainable AI (XAI) should have its utility tailored to the type of user and mission context (e.g., clear, straightforward information for quick decisions) and include more detailed information when the user needs to ensure that things are processed correctly.

What are the state-of-the-art measurements?

Behavioral

User errors; task completion

Surveys

Some InfoQ measures exist, but they are not well suited to combat systems.

New scales should be developed and validated for this purpose.

Qualitative

Analysis of Information Utilization; user interviews



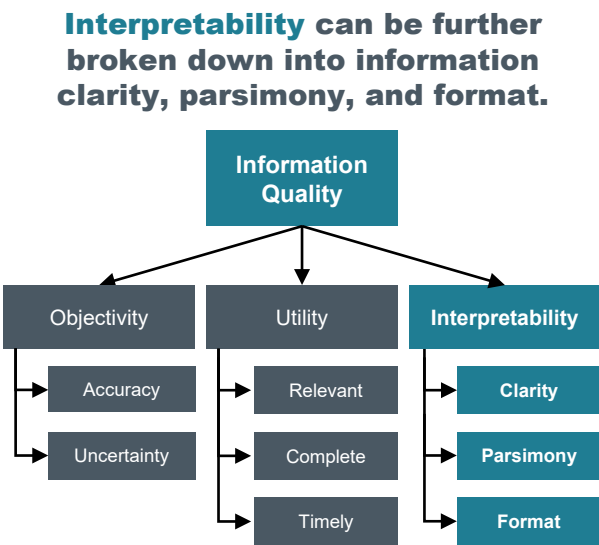
Information Quality: Interpretability

Information quality is determined by *objectivity*, *utility*, and *interpretability*. Interpretability focuses on whether information is communicated in a way the operator can understand. This is the focus of much of UI/UX work.

How is it relevant to testing?

The principles to achieve interpretable information are simple. The message must be clear; e.g., verbal or written communication should avoid jargon, difficult-to-remember concepts and ambiguous words. Different modalities have different best practices. Clarity is essentially the message's signal strength, with strong signals requiring less receiver sensitivity. Parsimony is about limiting the quantity of information and noise provided. Technically minded people often want to include all relevant information, but increasing the total quantity of information can often overload an operator's processing capacity. Finally, information needs to be conveyed in an appropriate format for its purpose.

TESSs should consider interpretability across the lifecycle; in particular, OT measurement should be under realistic workload spikes.



What could go wrong?

A multi-spectral data fusion device for checkpoints includes pop-up labels on every object it can identify—not only for weapons and bombs but also combs, chapstick, and wallets. This forces the operator to take on a larger task of sifting through all the data.

How can AI make it harder?

Conveying information in messages that are concise while still containing relevant information is a challenge for all systems. AIECs may have world models that are far enough removed from human understanding that it will be difficult to capture relevant information in short, relevant messages. One of the challenges of Explainable AI is being able to create interpretable information out of system decision-making and model performance. The delivered information quality will need to be clear, parsimonious, and formatted in ways that allow operators, testers, commanders, and other stakeholders to make informed decisions.

What are the state-of-the-art measurements?

Behavioral	Surveys	Qualitative
Probe understanding under different workload levels; user errors; task completion	Some InfoQ measures exist, but they are not well suited to combat systems.	<u>Analysis of Information Utilization</u> ; user interviews
	New scales should be developed and validated for this purpose.	



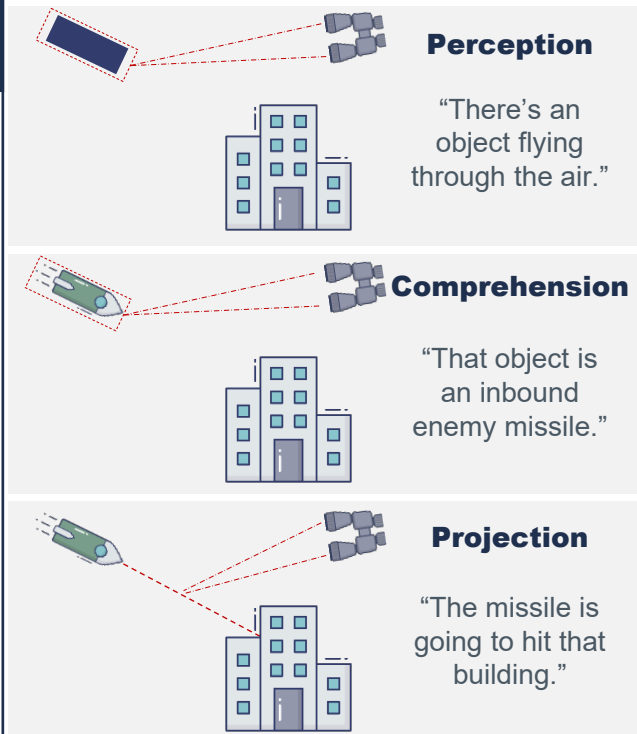
Situational Awareness

Situational awareness (SA) is the perception of the elements in the environment (including the internal state of the AIEC), the comprehension of their meaning, and the projection of their status in the future.

How is it relevant to testing?

SA is critical to making informed, timely decisions. Accessing and processing information are necessary but insufficient to have SA; operators must be able to consolidate data into a holistic understanding of the environment. Operators should be able to leverage their mental models, along with situational parameters, to forecast future events and dynamics. Operators do not need “perfect” SA to achieve mission success. Like evaluating trust, SA should be considered in context. Finally, an operator cannot be aware of something they do not know. Accordingly, SA should be evaluated relative to the ground truth and not be mistaken for perceived SA, which is an operator rating of their own SA.

SA is complex, and TESs should not commit to measuring SA without allocating adequate resources. Surveys measure only perceived SA and are insufficient alone.



What could go wrong?

A warfighter is overly dependent on an AIEC that identifies potential targets. They are aware of targets identified by the system but are unaware that the system has a high missed detection rate. The warfighter's poor SA of the internal state of their AIEC results in them missing a high-value target.

How can AI make it harder?

Complex automation expands the bounds of what an operator needs in order to have adequate SA of their current and future states. Beyond environmental and operational context, operators need SA of the AIEC's internal state. This is made more difficult by AIECs' internal states not being confined to finite rule-based logic that can be explicitly learned during training. Warfighter MMs will become critical to maintaining SA. As AIECs become more autonomous, mission performance will rely on warfighters and AIECs operating through a shared understanding of the problem state, and testing should examine to what extent agents have a common operating picture. Different understandings of the current situation will result in poorly coordinated decisions.

What are the state-of-the-art measurements?

Behavioral

Memory probes; reaction time; response to threats;

SAGAT; SPAM

Surveys

Self-report gives only perceived SA, not real SA. Not sufficient alone!

SART; SASHA

Qualitative

User interviews;

Goal-Directed Task Analysis
(GDTA)



Explainable AI (XAI)

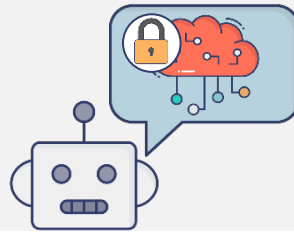
Explainable AI (XAI) here refers to active methods in which a system provides the “why” or the causal reasons for a system's internal logic and resulting output in a way that human operators can understand. Readers should note that there are competing definitions.

How is it relevant to testing?

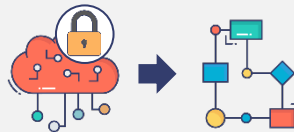
XAI is an evolving field that lacks a commonly agreed upon taxonomy. Regardless of framework, it is useful to distinguish between causal reasons provided in reference to specific actions by systems themselves (“explanations”) versus those that can be deduced by stakeholders (“traceability”), versus a general understanding of the system's decision process (“transparency”).

Explanations can be provided before, during, or after system action, and they can come in any modality that highlights the behavior's causality. Their “goodness” can be evaluated using information quality [1, 2, 3] metrics. Good XAI explanations are useful for deciding whether the AIEC performance will be consistent in future situations, which aids in developing mental models. For instance, knowing an AIEC failed trying to do the “right thing” provides feedback to the warfighter that is different from knowing it was pursuing a bad goal. Useful explanations must be tailored toward their intended audience.

TESSs should provide their definition of XAI and commit to measuring the effect of system explanations’ on mission performance and warfighter decision-making.



Many AIEC algorithms are too complex for humans to understand or are hidden in black boxes.



XAI methods try to translate the AIEC algorithms into something understandable by humans.



This translation must be communicated to the human in a way that meets **information quality** standards.

What are the state-of-the-art measurements?

There are currently no “off the shelf” solutions with widespread consensus to objectively characterize XAI.

What could go wrong?

Bad explanations can build incorrect mental models, and even well-implemented transparency can cause complacency. Both of these can lead to warfighters making errors they would not make without XAI.

How can AI make it harder?

Beyond traditional, automated decision-making tools that summarize, consolidate, and present information to end users, AIECs will be able to independently reach solutions and may eventually interact with the operator like a teammate. This may be particularly tricky, given that AIECs are often leveraged to improve performance on tasks, not to improve interactions with teammates. AIEC world models and decision models are typically built for tasks that humans are not well suited for, so those models may not be easily human understandable. In order for a human operator to calibrate their trust in and collaboratively problem solve with an AIEC, XAI methods must be developed to empower operators to understand the system's decision-making process.

HSI Concept One-Pagers

This subsection introduces 3 HSI concepts using the one-pager layout. These 3 concepts have been grouped together because they are most related to a warfighter's need to know when it is appropriate to leverage their technology.

**Warfighter
Need**

**"I need to be able to make good decisions
about where and how to use this system."**

**OODA Loop
Decision Stage**

Decide

**DoD Ethical
Principle**

Responsible

HSI Concepts

1. Trust and Reliance
2. Emergence
3. Workload





Trust and Reliance

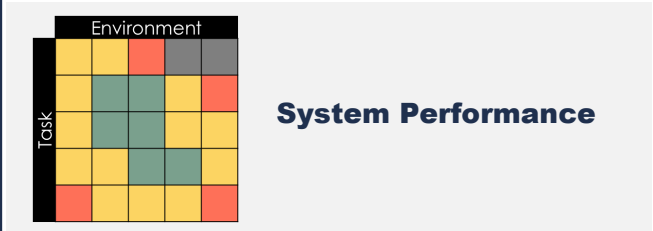
Trust is a person's belief that something can be depended on in vulnerable or uncertain situations. The critical behavioral outcome of trust is reliance, which is the use of the system in those situations.

How is it relevant to testing?

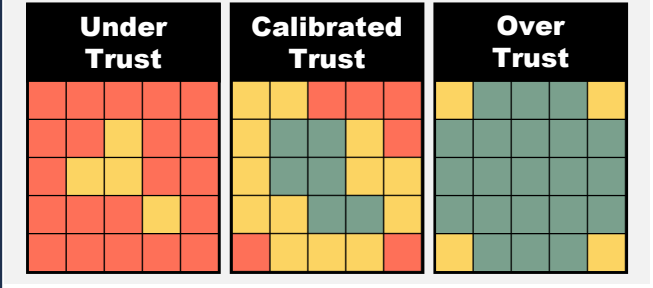
Both over- and under-trusting a system can lead to regrettable outcomes, so DoD's goal should be appropriately calibrating trust, not universally building it. Too much trust can endanger users who rely on it in conditions where the AIEC performs poorly. Conversely, too little trust may lead users to abandon use of the system when it could aid or protect them. Trust is task- and context-dependent, and it will evolve over time, so it must be measured accordingly. A single measure of holistic system trust is insufficient.

TESSs should commit to measuring warfighter trust across operational conditions and evaluating calibration relative to system performance. This should be done with new operators and those with field-representative experience levels.

A warfighter's trust is properly calibrated when their reliance on the system matches the system's performance.



Warfighter's reliance on the system



What could go wrong?

Soldiers had good experiences with their Optionally Manned Fighting Vehicle AI driver in the lowlands and have been letting it drive unsupervised. The operator allows the system to drive in mountainous terrain—where the AIEC has not been trained—and it ends up driving off a cliff.

How can AI make it harder?

User trust and system trustworthiness are often conflated, but knowing whether trust is calibrated requires knowing whether the system is trustworthy. The concept of trust in AIECs carries a lot of baggage, and people seem to expect either sci-fi miracles or the Terminator to appear by sprinkling “AI fairy dust.” Realistic expectations are uncommon. Civilian AI applications have seen over-confidence in unproven technology or near total rejection.

Trust is not a new issue, but it has not been a primary T&E concern. Current test strategies implicitly and explicitly require employing the system under test for test events, which makes evaluating reliance difficult.

What are the state-of-the-art measurements?

Behavioral	Surveys	Qualitative
Observed reliance; risk acceptance/taking	<u>Trust of Automated Systems Test (TOAST)</u> ; <u>Reliance Intentions Scale</u>	User interviews



Emergence

Behavior is considered emergent when the interaction of parts produces effects that the individual components do not have on their own. Emergent properties and behaviors can be any combination of expected/unexpected and desirable/undesirable.

How is it relevant to testing?

Most of the conversation around emergence in AI focuses on unexpected negative outcomes, but testers should be aware that this is an incomplete understanding. Expected and desired emergence, such as TTPs created to achieve mission goals, needs to be validated through T&E. This will be especially true when it comes to human-machine teaming. However, emergence also requires exploratory testing to looking for undesirable, unexpected behaviors.

Most people envision emergence as system-to-system or intra-system interactions that result in unusual behavior. Additionally, human operators can give unexpected or erroneous inputs that may cause the system to behave strangely or they may invent "off-label" uses for these technologies. Beyond considering how humans might result in unexpected AIEC behavior, testers must also consider how the system can change human behavior. Humans may execute their own tasks differently when working with AIECs.

TESSs should resource free-play testing where emergence can arise from all agents, and commit to following up on any unexpected emergent behavior observed in both structured and free-play testing.



A helicopter has an AIEC to identify enemy weapon emplacements to avoid hostile fires. The AIEC was designed to be risk-averse and has many false alarms.



The warfighters are unaware of the AIEC's risk-averse nature and invent a problematic "off-label use" for the AIEC, using the system to find targets.



What are the state-of-the-art measurements?

There are currently no "off the shelf" solutions to objectively characterize emergence. A future Part 2 of this framework will highlight promising research that should be further explored.

What could go wrong?

A helicopter is equipped with an AIEC to identify enemy weapon emplacements to avoid hostile fires. The AIEC was designed to be risk-averse and has a lot of false alarms. The warfighters are unaware of the AIEC's risk-averse nature and invent an "off-label use," using the system to identify targets.

How can AI make it harder?

The complexity and brittleness of AI makes emergent behavior much more likely, and warfighters will be expected to overcome system inadequacies on the fly. The main goals of emergence testing in DoD should be to confirm that expected, desirable, emergent properties or behaviors exist and are functioning, and to identify unexpected, undesirable emergence. If undesirable emergence is anticipated and subsequently accounted for through CONOPS, testers should assess the effectiveness of any procedures aimed at minimizing operational impact. To capture emergence that arises out of typical use not seen in operational testing (e.g., off-label use), testers will need to consider less controlled, more observational test designs.



Workload

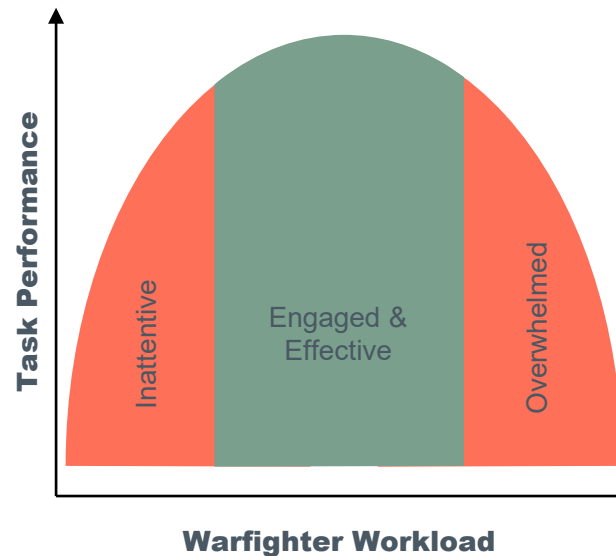
Workload comprises the physical, mental, and temporal resources required by the current tasks relative to the resources available to the person.

How is it relevant to testing?

Workload is a well known HSI concept and has a non-linear relationship to human task performance, where both high and low workload can be dangerous. Humans have limited task resources available, and greater task performance often requires greater effort. When those resources are exhausted (i.e., there are high workloads), people will begin shedding tasks and sacrificing performance on some tasks to maintain performance on others. High workloads hurt overall mission effectiveness and are dangerous if task failure has severe consequences. Low workload, however, is also problematic; humans suffer from inattention and may fail to react to changes in the environment.

TESSs should commit to measuring workload, in both nominal and off-nominal situations within safety constraints.

Human performance is optimal at moderate workloads



What could go wrong?

Like an overloaded air traffic controller, a warfighter overseeing multiple AIECs is given more autonomous systems to monitor than they can reasonably manage, and they miss a critical error. Conversely, another is given so little to do that they miss the critical warning when it finally occurs.

How can AI make it harder?

Many AI systems are intended to reduce operator workload or preserve task performance during task shedding, but they have been shown to do the opposite under certain circumstances, introducing “invisible work” and increasing operator workload instead. This should be tested. If humans are meant to oversee, monitor, or audit AI responses, both high and low workload can prevent human oversight from being meaningful.

What are the state-of-the-art measurements?

Behavioral

Reaction time; added-task performance change

Surveys

NASA Task Load Index;

AFFTC Revised Workload Estimate Scale;

Not Recommended: Bedford, Modified Cooper-Harper

Qualitative

Debriefs or after-action reviews

HSI Concept One-Pagers

This subsection introduces 3 HSI concepts using the one-pager layout. These 3 concepts have been grouped together because they are most related to a warfighter's need to meaningfully govern their technology to carry out their intent.

**Warfighter
Need**

**"I have to execute my decision. Make it easy to
get the system to do what I intend it to do."**

**OODA Loop
Decision Stage**

Act

**DoD Ethical
Principle**

Governability

HSI Concepts

1. Function Allocation
2. Usability
3. Training Quality



Function Allocation



Function allocation (FA) is the assignment of the collective work between human operators and their automated systems required to achieve mission goals.







How is it relevant to testing?

The allocation of work should be considered throughout CONOPS and system development; it must consider who is assigned a task (authority), who is accountable for a task (responsibility), and who is capable of completing a task (autonomy). Testers should evaluate whether the assigned authorities, responsibilities, and autonomies are suited for human and system capabilities and should identify limitations in order to help maximize mission performance.

Mission performance is usually degraded without a deliberate function allocation. Development has a tendency to automate as much as possible and leave a set of disjointed, difficult tasks to the operator. This poorly considered “leftover allocation” tends to cause workload lulls and spikes and adds “invisible work” with the system, such as communication, coordination, and oversight. For example, mismatches between an operator's assigned tasks and what outcomes they will be held responsible for can also lead to more invisible work, where the human feels obligated to monitor a system.

TESSs should require the PM to submit an FA for evaluation as part of the assurance case for the system.

Developers should include an FA, where all the work for a given task is allocated to the human or AIEC.

Task	Assigned
Identifying targets	 The AIEC has been given most tasks; however, the warfighter is responsible for the outcome and must terminate the engagement if something goes wrong.
Prioritizing targets	
Engaging targets	
Terminating engagement	
Monitoring	 This responsibility-authority mismatch has created “invisible” work not in the CONOPS.
Comm. & coordination	

What are the state-of-the-art measurements?

Testers should confirm that programs have a function allocation (i.e., identified who is assigned, capable, and responsible for each task the system will be used to perform).

Learn more [here](#) and [here](#).

What could go wrong?

A human driver responsible for an autonomous vehicle feels obligated to monitor it, even though they were not given supporting procedures or training. Because this work was not included in the CONOPS, the operator must also complete additional tasks and is often overwhelmed.

How can AI make it harder?

Testers should consider the allocation of work to confirm that all assigned tasks can be completed and that tasks can be appropriately traded (e.g., handoffs, interventions) between automation and humans. For example, if an AIEC decision cycle is faster than a human can monitor, humans cannot effectively govern the autonomy on-the-loop. Testers should consider observational designs to identify unacknowledged invisible work that arises out of the fielded use of a system. Finally, testers should confirm that operators are provided with appropriate training and procedures to successfully accomplish all their work.



Usability

Usability is the fitness of a tool for a task. It is composed of utility (whether the system has the capabilities one needs for a task) and ease of use (whether it is easy to get the system to do what one intends it to do).

How is it relevant to testing?

A perfectly usable system would be one that directly translates user intent into action. Utility is the extent to which the system is capable of contributing to that intent. If it does not have the right capabilities, it has low utility. Ease of use is the amount of effort required to accomplish that contribution. Usability is strongly related to both workload and training. More usable systems require less workload to complete the same task, essentially providing a buffer for when task demands spike during emergencies. Furthermore, more usable systems require less training to reach proficiency, which is especially critical in high-turnover positions. Information interpretability is related to usability.

For DT, TESs should evaluate usability at a granular subsystem level, whereas for OT TESs should holistically examine the systems.

Task: Test the missile warning system!

In order for the system to be usable for this task, it must have a test message functionality that is easy to use.



This system lacks a test capability, which makes it unusable for this task.



This system has a test button, but the buttons are close together and poorly labeled. This UI would score low for ease of use.

What could go wrong?

A warfighter intentionally aims to miss for a warning shot, but their aim assist cannot understand their intent and lacks a “warning shot” function. It “corrects” and kills the target.

How can AI make it harder?

Even in a system that autonomously executes its tasks, humans will still need to interact with it to give it its initial orders, extract additional information from it, or potentially intervene to alter or stop its behavior. Systems that execute actions, whether selected by the operator or by the system itself, needed to be assessed on whether those actions match the operator’s intent (e.g., think of autocorrect on your phone).

What are the state-of-the-art measurements?

Behavioral

Task completion; error rate; help desk tickets; automatic recording

Surveys

SUS; UMUX; UMUX-Lite

Qualitative

User interviews;

Heuristic walkthrough;

Cognitive walkthrough



Warfighter Training Quality

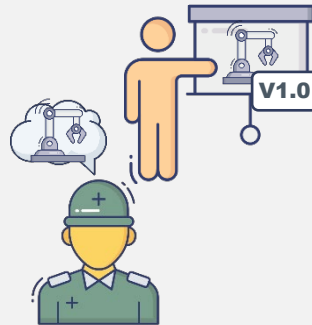
Training quality covers the extent to which the warfighter has been prepared to use a system during actual operations.

How is it relevant to testing?

Skill acquisition typically moves through three stages, starting as explicit knowledge that users consciously step through, and ultimately ending as relatively automatic and efficient processes. Early performance growth tends to be rapid, eventually leading to diminishing returns. A lot of formal equipment training focuses on providing operators with the first stage of explicit knowledge and rapid growth, while the slower (but important) skill maturing is left to informal training at the unit and “learning through doing.” When training is relevant, the fast rate of learning begins sooner, whereas poor training will require more trial-and-error before it accelerates.

TESs should commit to assessing training quality on representative operators, not engineers, contractors, or “golden” crews.

During Training



During training, the warfighter was taught how to use the extensive, complex functionalities of an AIEC, version 1.0.

1 year later



The warfighter has forgotten most of the complexity covered in training, and they have never been trained on the changes to the system.

What could go wrong?

The training on how a threat warning system makes its decisions is based on an outdated, prior version of the system that no longer matches how system decisions are made. Based on outdated knowledge, the operator believes the system is false-alarming when there is actually a real threat.

How can AI make it harder?

AI-enabled systems, especially those that learn continuously, will be less static than traditional systems. TTPs and CONOPS will co-evolve with system capabilities as we learn what is not technologically feasible, or what innovative “off-label” uses are invented for the system. What operators need to know can evolve quickly, meaning their training must be dynamic as well. T&E may need to focus not only on evaluating the quality of the training content but also on the process by which training will be updated and distributed. Finally, when AIECs take over tasks, it means that the user is no longer practicing those skills. This can lead to a loss of expertise for when the operator must intervene to prevent fatal accidents.

What are the state-of-the-art measurements?

Behavioral

Knowledge-based or applied-skill tests; benchmark comparisons over time

Surveys

Self-assessment of readiness is often inaccurate but may be feasible for some systems.

Qualitative

User interviews

Operational Assessment of Training Scale (OATS)

Summary of Recommended TES Actions

	HSI Concept	TESs Actions
Observe & Orient	Mental Models (MMs)	Assess warfighters' (WFs) MM and evaluate how well MMs allow WF to predict system behavior.
	Boundary Awareness	Evaluate WFs' knowledge of system limitations.
	Situational Awareness (SA)	Employ SA measures beyond self-report. TESs should not commit to this if adequate resources will not be assigned.
	Info Quality: Objectivity	Compare the accuracy and uncertainty of information provided versus WF needs across operational conditions.
	Info Quality: Utility	Test information utility with real WFs in both DT and OT.
	Info Quality: Interpretability	Measure under operationally realistic workload spikes in OT events.
	Explainable AI (XAI)	Identify which XAI definition you adopted for your test and measure system explanations and impact on WF decision-making.
Decide	Trust & Reliance	Measure WF trust across operational conditions and evaluate calibration relative to system performance.
	Emergence	Resource free-play testing where emergence can arise from all agents, and follow up on any emergent behavior.
	Workload	Measure nominal workload as well as off-nominal workload within safety constraints.
Act	Function Allocation (FA)	Require programs to submit an FA for evaluation as part of the assurance case for the system.
	Usability	Evaluate usability at a granular subsystem level for DT, and holistically examine the system-of-systems in OT.
	Training Quality	Assess training quality on representative WFs—not engineers, contractors, or “golden” crews.



HSI T&E over the AIEC lifecycle

This Section:

- + Highlights how traditional software-intensive systems are stretching current T&E processes to their limit, and how AIECs will exacerbate existing shortfalls and pose novel challenges
- + Advocates for integrating T&E continuously throughout the system lifecycle, including during acquisition (“shifted left”) and sustainment (“shifted right”)

04



We cannot keep testing and evaluating HSI the same way we have been

As we incorporate T&E throughout the system lifecycle, where does HSI fit in?

Our current processes are too siloed and static to support T&E of AIECs.

Traditional systems are already stretching our current T&E processes to their limit. AIECs promise to exacerbate existing shortfalls and pose novel T&E challenges. These challenges will not only pose problems for the T&E of system performance but also the T&E of HSI. In many cases, DoD already fails to characterize the quality of HSI for traditional systems.

AIECs signal a need for a paradigm shift. We must integrate T&E continuously throughout the system lifecycle, including during both acquisition and sustainment. Doing so requires resources, including greater access to operators and data during development and operations, onboard instrumentation and testbeds, and access to HSI expertise.

T&E of HSI must be integrated throughout the lifecycle of AIECs.

AIECs are changing the way warfighters achieve their mission objectives. AIECs are performing more complex tasks and, in some cases, may behave like teammates. Therefore, it is critical to understand early in the design process how effectively operators interact with the AIEC so that any necessary changes can be made before the design is finalized (i.e., we must shift left).

Additionally, as both CONOPS and the AIEC evolve, the quality of HSI may change over time, requiring post-deployment T&E to identify drift and mitigate negative consequences or provide support for new developments (i.e., we must also shift right).

T&E methods must evolve to account for the challenges that AIECs impose.

For traditional systems, historical information can help predict and scope the cost of failure modes. The advanced and complex interactions of AIECs, however, are hard to characterize. To ensure appropriate, timely T&E at a reasonable cost, we need to acquire systems that are built to facilitate T&E, develop testbeds and instrumentation, and focus on building a body of evidence for system performance over time. In order to generalize our test results, we must develop methods for obtaining and validating causal models of AIEC decision-making and behavior. Without these steps, characterizing the effects and understanding the causes of AI influence on mission performance, including HSI, will likely be an insurmountable challenge.



We must continue to “shift left”

An ounce of prevention is worth a pound of cure.

If you are going to engineer with humans in mind, it is easiest to start early, when it's easiest to shape design.

What might seem like a sensible design choice to an engineer may not make sense for operational users. Warfighters need to be involved throughout the acquisition process to unearth the unexpected as early as possible.

Testers can play a key role in shifting left by working with operational users to ensure that requirements are both testable and operationally relevant. Testers should insist that operational users interact with the system iteratively throughout development in order to promote effective interaction with the AIEC and appropriately calibrate trust.



Shift Left



Function Allocation

The assignment of the collective taskwork between human and automated agents required to achieve mission goals.

HSI design choices should be explicitly articulated and tested early and often

AIEC design choices have downstream consequences for HSI quality.

A warfighter's interaction with technology is not limited to the interface, particularly for AIECs. A complex, unpredictable system cannot be transformed into an understandable, trustworthy system solely through interface design. Systems' decision processes impact warfighters' ability to predict system behavior and meaningfully engage with the task. Additionally, many algorithms do not translate well to human-understandable explanations.

T&E must consider the implications of how algorithm selection will impact warfighters' ability to leverage technology to achieve mission success. Testers should get involved in T&E of candidate algorithms, putting algorithms in front of operational users early and often to ensure that the technology is usable and understandable.

Poorly defined CONOPS and function allocation are difficult to fix later in the acquisition.

Many HSI problems can be avoided by understanding the work domain at the beginning of the design process. All work should be explicitly allocated following best practices.

The PM should provide resources for a function allocation assessment in the TES and make the results of this assessment available to testers as part of the body of evidence required to assess AIEC performance and to aid in later test planning.

Testers should design tests to determine whether taskwork is allocated in a way that degrades mission performance. These tests should cover a range of operational scenarios to include those that are routine and those that are highly stressful. These tests should be resourced in the TES.



Shift Left



User Touchpoint

An event at which the people who will use the system are brought into the development process.

Relevant stakeholders should interact with programs early and often

Stakeholders should be integrated throughout the acquisition process.

The full breadth of relevant stakeholders and uses of the system should be considered throughout the development process. Relevant stakeholders include operational users and maintainers and other possible users, such as the administrators and commanders who will make decisions on when and where the system will be deployed.

Early interactions with operational stakeholders are necessary to ensure that the appropriate CONOPS, requirements, and system architectures are in place to provide capabilities that will empower warfighters to maintain (and ideally improve) mission performance. Early interactions also promote a robust and well-resourced test strategy that produces T&E information that can be leveraged to make informed program decisions.

User interactions can serve a variety of purposes and should be tailored to system maturity.

User touchpoint events can serve a variety of purposes and can be tailored to where the program is in the development process. Early interactions with users can help develop hypothetical CONOPS and mock-up user interfaces, and can help ensure that the system is developing capabilities that help users accomplish their missions. Later touchpoints can encourage users to interact with prototypes for system refinement, build training procedures for confusing aspects of the system, and get feedback on how performance can meet expectations.

Developers often include touchpoints in development, and testers should collect data from or run these events themselves to add to the body of evidence. These touchpoints should be identified and resourced in the TES.



We must continue to “shift right”

All AIECs are software-intensive systems that will continue to receive updates over time. This means that their capabilities, functionality, and interaction with users may also change over time. These changes may improve or degrade human-system performance.

Given the expansive state-space of AIECs, it is unreasonable to test all functions under all contextually relevant environments before deployment. DoD will need to monitor AIECs in the field to ensure that they are behaving within expected operating parameters. Continued data collection must be planned for and resourced, and PMs must make this data available to testers.

T&E cannot stop at deployment.

We need a post-fielding TES.



Emergent behaviors should be monitored throughout deployment

Emergent behaviors will happen.

Not all emergent behaviors are undesirable; indeed, DoD should confirm that expected, desirable emergent behaviors exist during operational testing. However, unexpected emergent behaviors introduce problems to the current T&E model, because there is no way to guarantee that all unanticipated behaviors will arise during testing.

As fielded AIECs interact with each other during operations, cases of emergent behavior may increase. T&E must shift right to capture unexpected emergent behavior, and identify unacceptable performance degradation that should be addressed. The TES should address monitoring for emergent behaviors and required resources.

Warfighters are often overlooked sources of emergent behavior.

Humans are creative and often leverage tools they are given to perform “off label” tasks, such as using a screwdriver to pry up staples. However, with the complexity of AIECs, unintended use could result in unpredictable outcomes.

T&E needs to identify the intended use of AIECs, then collect data in both operational testing and fielded operations to ensure that employment remains within expected parameters. As new, off-label uses arise, they should be documented and programs should decide what, if anything, should be done. This could include updating CONOPS, changing documentation or training, or creating new limitations on how the system can be used.

Shift Right



Emergent Behavior

Behavior stemming from the interaction of parts that the individual components cannot carry out in isolation.



Important!

Identify & document “Shift Right” activities early and often!



Warfighters learn in the field; evolving expertise and training should be re-evaluated

As systems evolve and perform increasingly complex tasks, warfighters will change too.

Continuous learning presents the possibility that systems may require re-certification after deployment. However—less often acknowledged—is that warfighters will evolve as well. As AIECs take over tasks, warfighters often lose engagement with their work. In adapting to new roles, such as monitoring, warfighters lose or simply never develop pertinent knowledge and manual skillsets. However, during off-nominal operations that require intervention, this lost expertise is critical to mission success. T&E must consider post-fielding evaluations of not just the AIEC but also of warfighters' evolving expertise and mental models.

The TES should address this possibility and indicate how and when warfighters' evolving expertise and mental models will be evaluated.

Building effective training programs will become more difficult.

Training warfighters to operate systems is a critical task that is made more difficult by the complexity of AIECs and their typical lack of transparency. Training programs must evolve and adapt over time to account for unpredictable emergent behaviors and AIECs that learn. This adaptation will likely be exacerbated by the steep learning curve and the high skill levels required to understand how to work with these novel systems.

A static evaluation of training quality before deployment will not be representative of the training's effectiveness once the AIEC and warfighter evolve. As both systems and their accompanying training programs change, we must reassess the efficacy of these programs.

Shift Right



Training Quality

Training quality covers the extent to which the warfighter has been prepared to use a system during actual operations.



Important!

Identify & document "Shift Right" activities early and often!



Test populations must be representative and given opportunities for “free play”

T&E must identify critical gaps in warfighter mental models and AIEC behavior.

T&E must consider whether the warfighters in our test are representative of warfighters in the field. Consider a test performed with operators of a legacy system. These evaluations would not be sufficient to characterize the performance of a novice warfighter. Furthermore, to capture how warfighter mental models evolve over time with experience, testing will need to continue post-fielding.

Additionally, no warfighter will be able to fully predict a system's behavior. Testing must identify performance- or safety-critical mental model gaps to determine whether they should be addressed with training or design changes. Evaluations of mental model gaps should be addressed and resourced in the TES.

T&E should plan to discover novel, unintentional, or undesirable uses or outcomes.

As our warfighters adapt to working with AIECs, shifting right will be crucial to accurately characterizing the performance and reliability of these systems once fielded. As previously discussed, warfighters will evolve and use tools in unanticipated ways. Free-play events will also be critical for understanding trust and reliance. T&E is more likely to capture these deviations when:

- Warfighters have the freedom to use the system in new ways.
- Operators have enough time to “invent” new uses. They are unlikely to make up new uses on day one.
- T&E monitors the system’s usage outside of standard operating conditions.

Shift Right



Mental Model

A warfighter’s mental model is their set of knowledge. When applied to automation, these models allow them to infer the current state of a system and anticipate future states.



Important!

Identify & document “Shift Right” activities early and often!

Dependable human interaction with AIECs is necessary to achieve success across domains

Testers can leverage HSI T&E to link traditionally siloed evaluations and begin integrating evaluation efforts.

HSI is essential for fielding trustworthy AIECs beyond mission effectiveness.

This document focuses on mission performance; however, the same HSI principles critical to effectiveness apply to other domains, such as safety, suitability, cybersecurity, and ethical employment. All of these domains have different contributions to mission success, but they share a commonality: end users.

Regardless of whether a warfighter needs to safely, securely, or ethically employ an AIEC, they must be able to execute the OODA loop: (1) understanding and predicting the situation, (2) making decisions about when to use a tool, and (3) knowing how to govern a tool (discussed in [Section 02](#)).

AIECs increase the need to break down stove-piping between evaluation efforts.

Authority over these different elements lies with separate organizations, and different working groups are generating independent solutions to what has traditionally been their own silo. However, AI can increase the interdependency of these different elements to an even greater extent.

Given the large burden of evidence demanded by AIECs and the difficulty of assembling an assurance case for them, it will be critical to not just deconflict our T&E efforts but also to ensure that they are mutually reinforcing. Given that human warfighters are the shared components across domains, HSI is a low-hanging fruit to begin integrating the existing lines of effort addressing novel AIEC challenges.

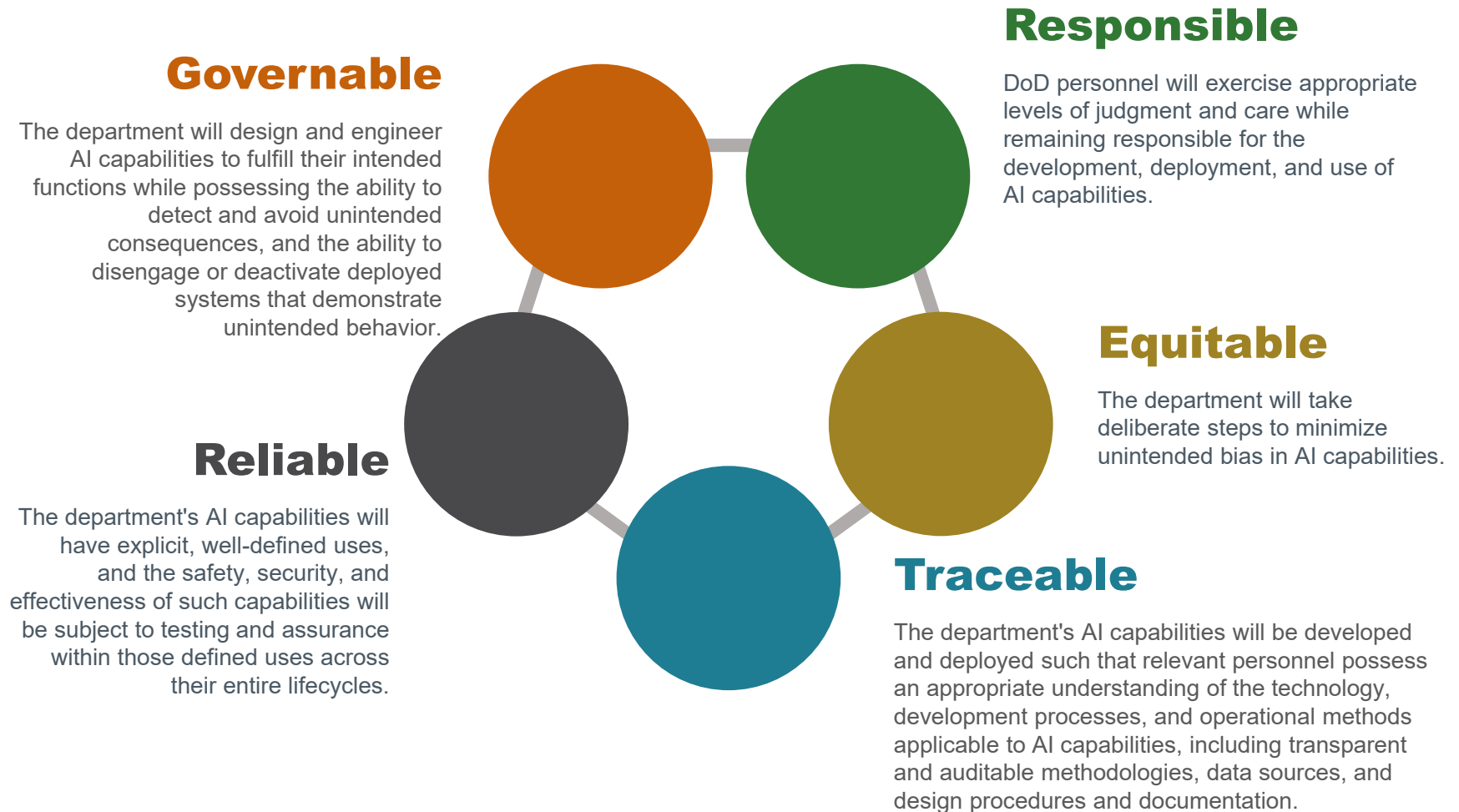
Methods and subject matter expertise for the T&E of ethical employment is in its infancy.

Since DoD proposed their five Ethical Principles in early 2020, there has been an open question of how to best operationally define and evaluate these principles. While some principles (e.g., “Reliable”) have obvious connections to traditional evaluations, others do not. Evaluations of ethical employment will be multifaceted and require disparate sources of assurance, but HSI provides a way to start operationalizing these high-level principles. Even though HSI provides a way to begin integrating all of the evaluation domains, we focus on ethical employment in this section, as it remains the least resolved of these issues.



Quick Reference to the DoD's Five Principles of Artificial Intelligence Ethics

In 2020, DoD adopted principles regarding the ethical use and development of AI systems. The next step is operationalizing these principles for different stakeholders.



Dependable human interaction is necessary for both effective and ethical employment

In the hunt for how to best operationalize and evaluate DoD's Ethical Principles, HSI characterizations are a critical piece of the puzzle.

For an AIEC to be **traceable**, warfighters must be able to understand how it works, predict how it reacts to inputs, and access relevant information.

To **responsibly** employ an AIEC, the warfighter must be able to make informed, timely decisions about when it is appropriate to use the system.

For an AIEC to be **governable**, the warfighter must be able to easily operate the system, including terminating the AIEC should the need arise.

DoD Ethical Principles

Responsible

Equitable

Traceable

Reliable

Governable

Warfighter needs mapped to decision stages of the OODA loop

**Observe
& Orient**

"I have to understand and predict the situation. Tell me what I need to know, when I need to know it, in a way that I understand."

Decide

"I need to be able to make good decisions about where and how to use this system."

Act

"I have to execute my decision. Make it easy to get the system to do what I intend it to do."



Reflecting on HSI T&E of AIECs

This Section:

- + Discusses how successful HSI T&E is critical for deploying trustworthy AIECs
- + Illustrates how human interaction is foundational to mission success across domains and across a capability's lifecycle
- + Explains why core HSI concepts should be explicitly addressed and resourced in TESS

05



Framework Recommendations



Successful HSI T&E is critical for deploying trustworthy AIECs.

In order to design useful and adequate system evaluation strategies, testers must actively consider where and how AI-enabled systems are incorporated into operators' task execution. The OODA loop is a familiar topic that can help testers frame this problem.



Human interaction is foundational to mission success across domains.

HSI bridges effectiveness, safety, cybersecurity, and ethics, and programs should emphasize HSI and its evaluation across systems' cradle-to-grave lifecycles. This will require organizational restructuring and resource commitment both early and often, throughout development and likely further into sustainment.



Core HSI concepts should be explicitly addressed and resourced in TESSs.

Testers should ensure that TESSs commit to triangulating HSI concerns through a combination of behavioral, survey, and qualitative methods. TESSs should make sure that HSI will be measured so that all measurement modalities can be tied to specific test points and outcomes.



Useful Free Resources Available Online

Below is an inexhaustive set of useful resources related to AI T&E and HSI measurement. Sources are linked or can be googled under that name.

Source	Benefit
<u>IDA Trustworthy Autonomy: A Roadmap to Assurance</u>	In-depth discussion of T&E challenges of AI-enabled or autonomous systems as well as possible solutions. Contains a section on HSI.
<u>MeasuringU.com</u> Blogs	Layperson-friendly explanations of usability testing tools and techniques.
IDA's <u>TestScience.org</u>	Tools and tutorials on many T&E issues.
<u>DOT&E Validated Scale Repository</u>	Scales and scoring guides for validated HSI scales recommended by DOT&E. Scales not available for every concept.
<u>Eurocontrol Human Performance Repository</u>	Search and sidebar contain a variety of HSI measurement topics, with tools and explanations provided for these issues and methods.
<u>MITRE HMT Systems Engineering Guide</u>	Guide on HSI development processes; also describes many formal qualitative analysis techniques. Many are not appropriate for T&E, but some can be adapted if programs need them.
<u>UsabilityBok.org</u> Methods	Usability design and evaluation resources that include explanations and links to the relevant literature.
<u>Nielson Norman Group</u>	Videos and blogs explaining methods produced by an UI/UX-focused research group.

