U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND ARMAMENTS CENTER

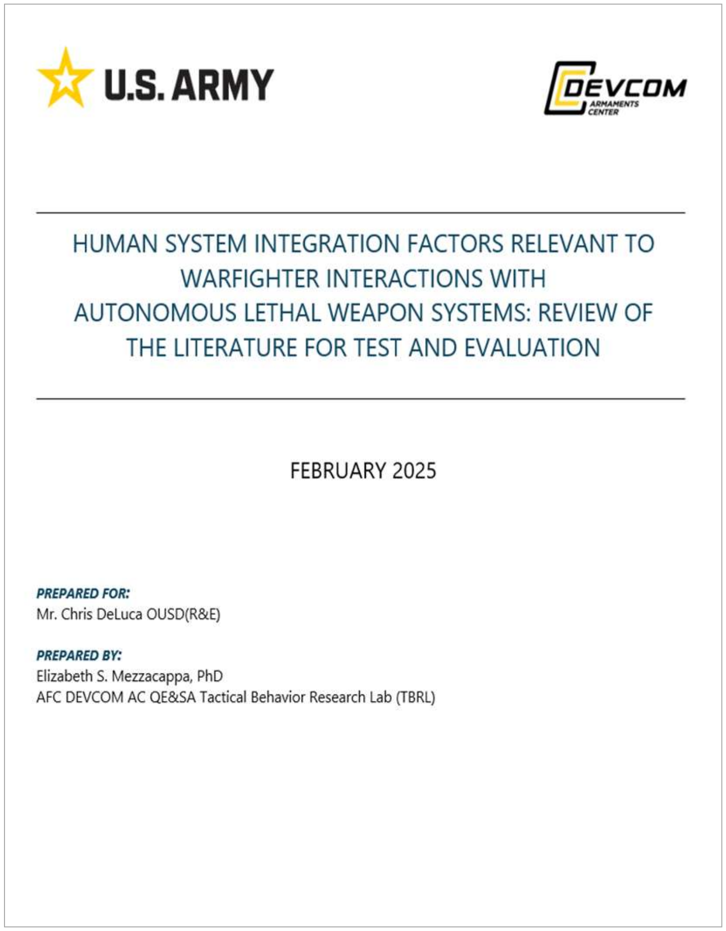
Appropriate levels of human judgement for autonomy
Presented at the 2025 AI4SE & SE4AI Workshop

17 SEP 2025

COMBAT CAPABILITIES DEVELOPMENT COMMAND | DR. ELIZABETH MEZZACAPPA, TACTICAL BEHAVIOR RESEARCH LABORATORY

- US Army Combat Capabilities Command Armament Center Mock Weapon Review

- Human Systems Integration Toolkit for DoDD 3000.09



U.S. ARMY

DEVCOM ARMAMENTS CENTER

HUMAN SYSTEM INTEGRATION FACTORS RELEVANT TO WARFIGHTER INTERACTIONS WITH AUTONOMOUS LETHAL WEAPON SYSTEMS: REVIEW OF THE LITERATURE FOR TEST AND EVALUATION

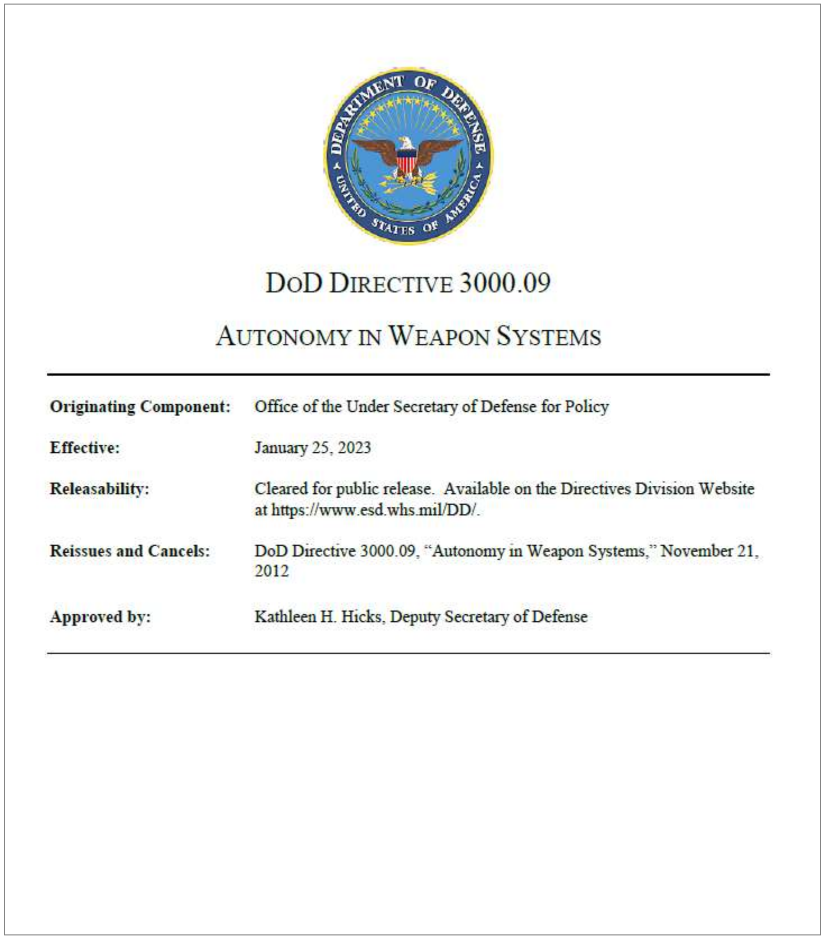
FEBRUARY 2025

PREPARED FOR:
Mr. Chris DeLuca OUSD(R&E)

PREPARED BY:
Elizabeth S. Mezzacappa, PhD
AFC DEVCOM AC QE&SA Tactical Behavior Research Lab (TBRL)

OUTLINE

- Extracts from DoDD 3000.09 that address appropriate levels of human judgement and testing and evaluation
- The Human Systems Integration (HSI) approach to appropriate levels of human judgement
- Prior work by Institute for Defense Analyses, (IDA) Chief Digital and AI Office (CDAO) and MITRE
- Overview of existing instruments
- Specific guidance from DoDD 3000.09 on testing and evaluation related to HSI
- Present a grouping these specifics into broader HSI topic areas
- Initial recommendations for testing and evaluation (T&E) of broad groups of human systems variables
- Tie back testing and evaluation of appropriate levels of human judgement



DoD Directive 3000.09

Autonomy in Weapon Systems

Originating Component:	Office of the Under Secretary of Defense for Policy
Effective:	January 25, 2023
Releasability:	Cleared for public release. Available on the Directives Division Website at https://www.esd.whs.mil/DD/.
Reissues and Cancels:	DoD Directive 3000.09, "Autonomy in Weapon Systems," November 21, 2012
Approved by:	Kathleen H. Hicks, Deputy Secretary of Defense

DODD 3000.09 AUTONOMY IN WEAPON SYSTEMS

1.2. POLICY.

a. Autonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.

G.2. DEFINITIONS.

Unless otherwise noted, these terms and their definitions are for the purpose of this directive.

TERM	DEFINITION
autonomous weapon system	A weapon system that, once activated, can select and engage targets without further intervention by an operator. This includes, but is not limited to, operator-supervised autonomous weapon systems that are designed to allow operators to override operation of the weapon system, but can select and engage targets without further operator input after activation.

DODD 3000.09 AUTONOMY IN WEAPON SYSTEMS (CONT'D)

(1) Systems will go through rigorous hardware and software verification and validation (V&V) and realistic system developmental and operational test and evaluation (T&E) in accordance with Section 3.

SECTION 3: VERIFICATION AND VALIDATION AND TESTING AND EVALUATION OF AUTONOMOUS AND SEMI-AUTONOMOUS WEAPON SYSTEMS

Regardless of the acquisition pathway or OSD T&E oversight status for a given weapon system, to ensure autonomous and semi-autonomous weapon systems function as anticipated in realistic operational environments against adaptive adversaries and are sufficiently robust to minimize failures:

a. Systems will go through rigorous hardware and software V&V and realistic system developmental and operational T&E, including analysis of unanticipated emergent behavior.

HOW DO WE TEST FOR “APPROPRIATE LEVELS OF HUMAN JUDGEMENT”?

DODD 3000.09 AUTONOMY IN WEAPON SYSTEMS (CONT'D)

(2) Consistent with the potential consequences of an unintended engagement or unauthorized parties interfering with the operation of the system, physical hardware and software will be designed with appropriate:

(a) System safety, anti-tamper mechanisms, and cybersecurity in accordance with DoD Instruction (DoDI) 8500.01 and Military Standard 882E.

(b) Human-machine interfaces and controls.

(c) Technologies and data sources that are transparent to, auditable by, and explainable by relevant personnel.

(3) For operators to make informed and appropriate decisions regarding the engagement of targets, the human-machine interface for autonomous and semi-autonomous weapon systems will:

(a) Be readily understandable to trained operators, such as by clearly indicating what actions operators need to perform and which actions the system will perform.

(b) Provide transparent feedback on system status.

(c) Provide clear procedures for trained operators to activate and deactivate system functions.

DODD 3000.09 AUTONOMY IN WEAPON SYSTEMS (CONT’D)

(3) V&V and T&E:

(a) Assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions, including possible adversary actions, consistent with the potential consequences of unintended engagement or unauthorized parties interfering with the operation of the system.

(b) Have demonstrated that the system can be reprogrammed with sufficient rapidity to enable timely correction of any unintended system behaviors that may be observed or discovered during future system operations.

(4) Adequate training, TTPs, and doctrine are available, periodically reviewed, and used by system operators and commanders to understand the functioning, capabilities, and limitations of the system’s autonomy in realistic operational conditions.

(5) System design and human-machine interfaces are readily understandable to trained operators, provide transparent feedback on system status, and provide clear procedures for trained operators to activate and deactivate system functions.

(6) For systems incorporating AI capabilities, the deployment and use of the AI capabilities in the weapon system will be consistent with the DoD AI Ethical Principles and the DoD Responsible AI Strategy and Implementation Pathway.

DODD 3000.09 AUTONOMY IN WEAPON SYSTEMS (CONT'D)

SECTION 4: GUIDELINES FOR REVIEW OF CERTAIN AUTONOMOUS WEAPON SYSTEMS

c. **Before a decision to enter formal development**, the USD(P), USD(R&E), and VCJCS will verify that:

(1) The system design incorporates the necessary capabilities to allow commanders and operators to exercise appropriate levels of human judgment over the use of force in the envisioned planning and employment processes for the weapon.

d. **Before fielding**, the USD(P), USD(A&S), and VCJCS will verify that:

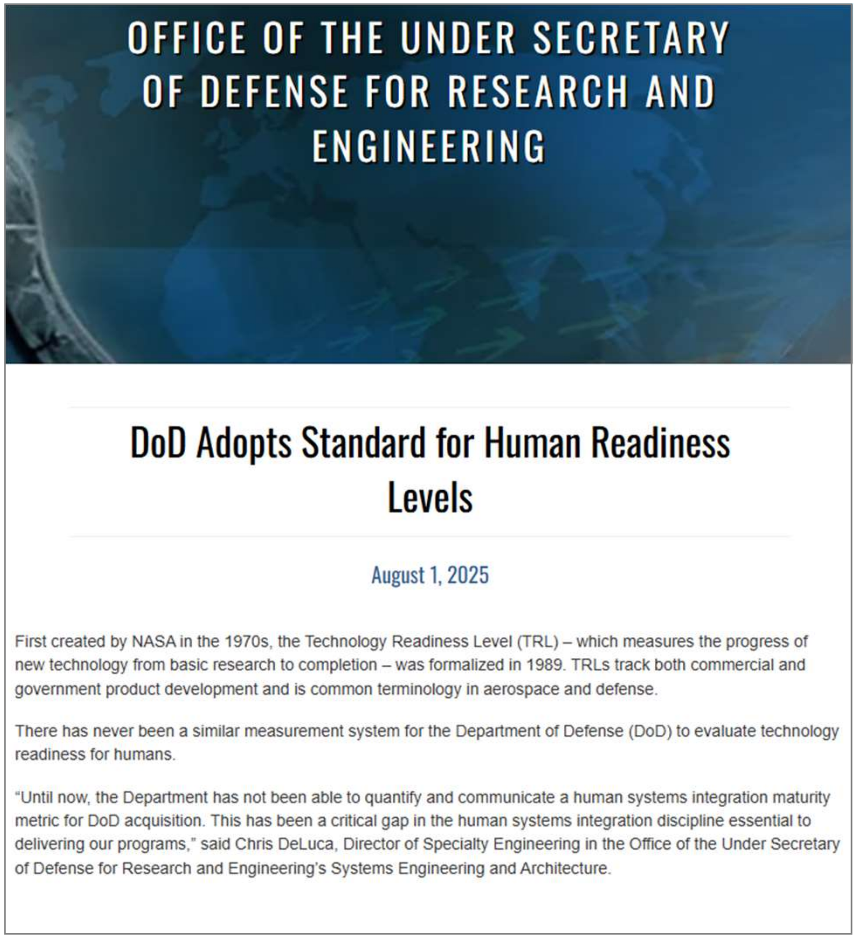
(1) System capabilities, human-machine interfaces, doctrine, TTPs, and training have been demonstrated to allow commanders and operators to exercise appropriate levels of human judgment over the use of force and to employ systems with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and ROE that are applicable or reasonably expected to be applicable.

HOW DO WE TEST FOR “APPROPRIATE LEVELS OF HUMAN JUDGEMENT?”

HOW DO WE TEST FOR “APPROPRIATE LEVELS OF HUMAN JUDGEMENT?”

Test for appropriate levels of Human Systems Integration.

HUMAN READINESS LEVELS



OFFICE OF THE UNDER SECRETARY OF DEFENSE FOR RESEARCH AND ENGINEERING

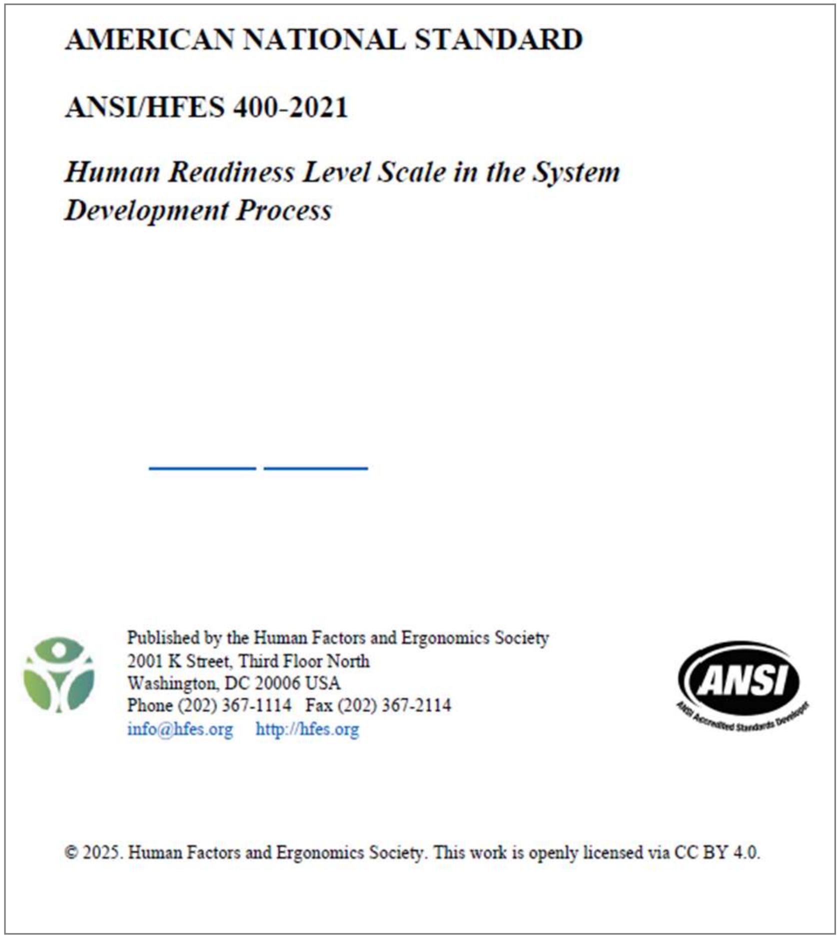
DoD Adopts Standard for Human Readiness Levels

August 1, 2025

First created by NASA in the 1970s, the Technology Readiness Level (TRL) – which measures the progress of new technology from basic research to completion – was formalized in 1989. TRLs track both commercial and government product development and is common terminology in aerospace and defense.

There has never been a similar measurement system for the Department of Defense (DoD) to evaluate technology readiness for humans.

"Until now, the Department has not been able to quantify and communicate a human systems integration maturity metric for DoD acquisition. This has been a critical gap in the human systems integration discipline essential to delivering our programs," said Chris DeLuca, Director of Specialty Engineering in the Office of the Under Secretary of Defense for Research and Engineering's Systems Engineering and Architecture.



AMERICAN NATIONAL STANDARD

ANSI/HFES 400-2021

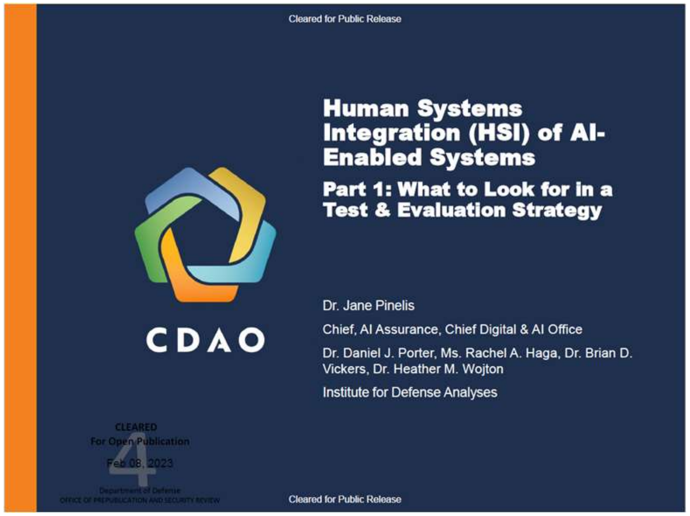
Human Readiness Level Scale in the System Development Process

Published by the Human Factors and Ergonomics Society
2001 K Street, Third Floor North
Washington, DC 20006 USA
Phone (202) 367-1114 Fax (202) 367-2114
info@hfes.org http://hfes.org

TABLES

DODD 3000.09 HUMAN SYSTEMS ENGINEERING REQUIREMENTS	
Describe how the system supports appropriate levels of human judgement.	Explain Soldier-Centered design history, incorporation of subject matter expert guidance and Soldier testing that demonstrates system support for appropriate levels of human judgement. Identify the human factors principles and Soldier testing data incorporated into design.

HUMAN SYSTEMS INTEGRATION FOR DODD 3000.09: CHIEF DIGITAL AND AI OFFICE



Cleared for Public Release

Human Systems Integration (HSI) of AI-Enabled Systems

Part 1: What to Look for in a Test & Evaluation Strategy

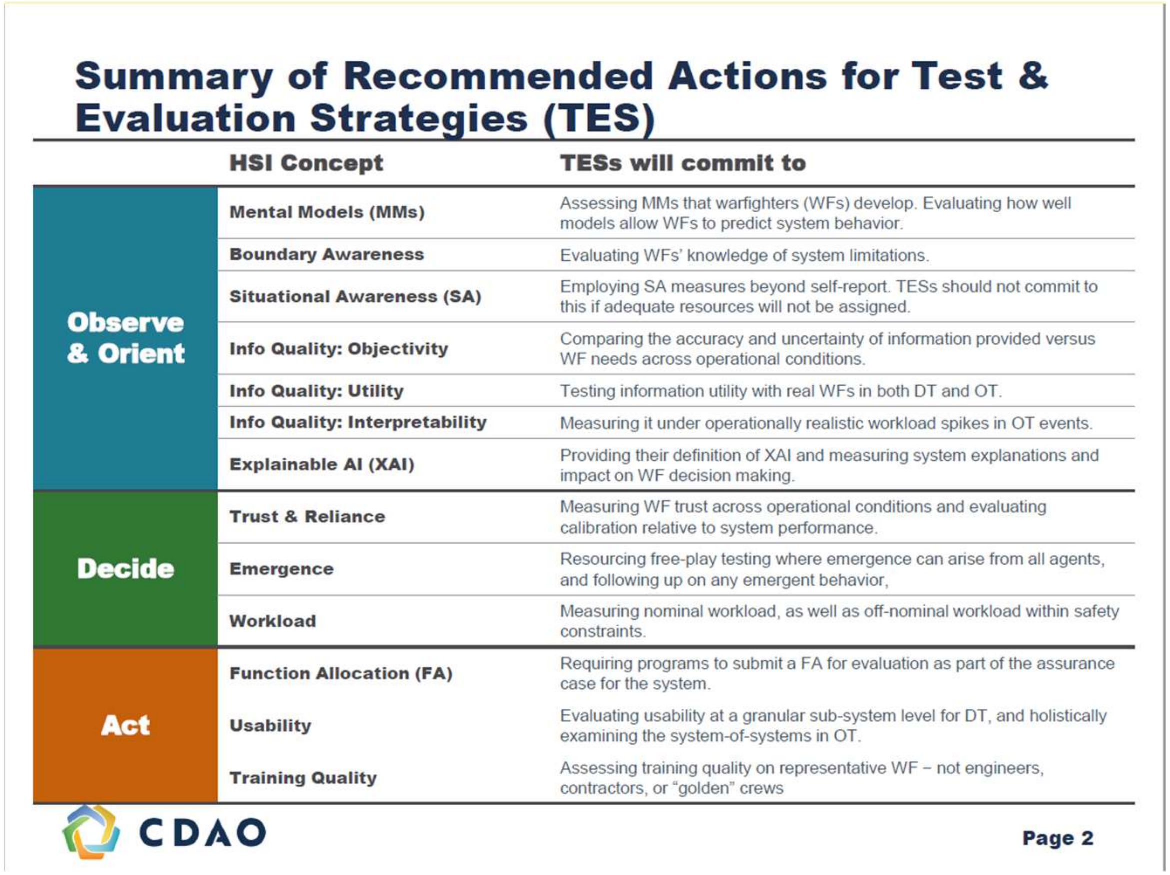
Dr. Jane Pinelis

Chief, AI Assurance, Chief Digital & AI Office

Dr. Daniel J. Porter, Ms. Rachel A. Haga, Dr. Brian D. Vickers, Dr. Heather M. Wojton

Institute for Defense Analyses

Cleared for Public Release

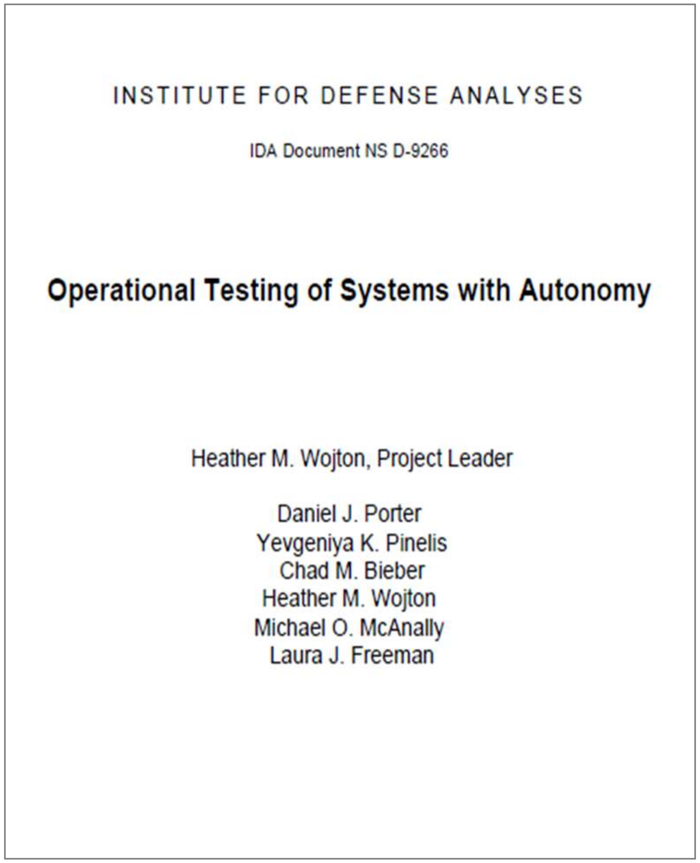


Summary of Recommended Actions for Test & Evaluation Strategies (TES)

	HSI Concept	TESs will commit to
Observe & Orient	**Mental Models (MMs)**	Assessing MMs that warfighters (WFs) develop. Evaluating how well models allow WFs to predict system behavior.
	Boundary Awareness	Evaluating WFs' knowledge of system limitations.
	Situational Awareness (SA)	Employing SA measures beyond self-report. TESs should not commit to this if adequate resources will not be assigned.
	Info Quality: Objectivity	Comparing the accuracy and uncertainty of information provided versus WF needs across operational conditions.
	Info Quality: Utility	Testing information utility with real WFs in both DT and OT.
	Info Quality: Interpretability	Measuring it under operationally realistic workload spikes in OT events.
	Explainable AI (XAI)	Providing their definition of XAI and measuring system explanations and impact on WF decision making.
Decide	**Trust & Reliance**	Measuring WF trust across operational conditions and evaluating calibration relative to system performance.
	Emergence	Resourcing free-play testing where emergence can arise from all agents, and following up on any emergent behavior,
	Workload	Measuring nominal workload, as well as off-nominal workload within safety constraints.
Act	**Function Allocation (FA)**	Requiring programs to submit a FA for evaluation as part of the assurance case for the system.
	Usability	Evaluating usability at a granular sub-system level for DT, and holistically examining the system-of-systems in OT.
	Training Quality	Assessing training quality on representative WF – not engineers, contractors, or "golden" crews

Page 2

HUMAN SYSTEMS INTEGRATION FOR DODD 3000.09: INSTITUTE FOR DEFENSE ANALYSES



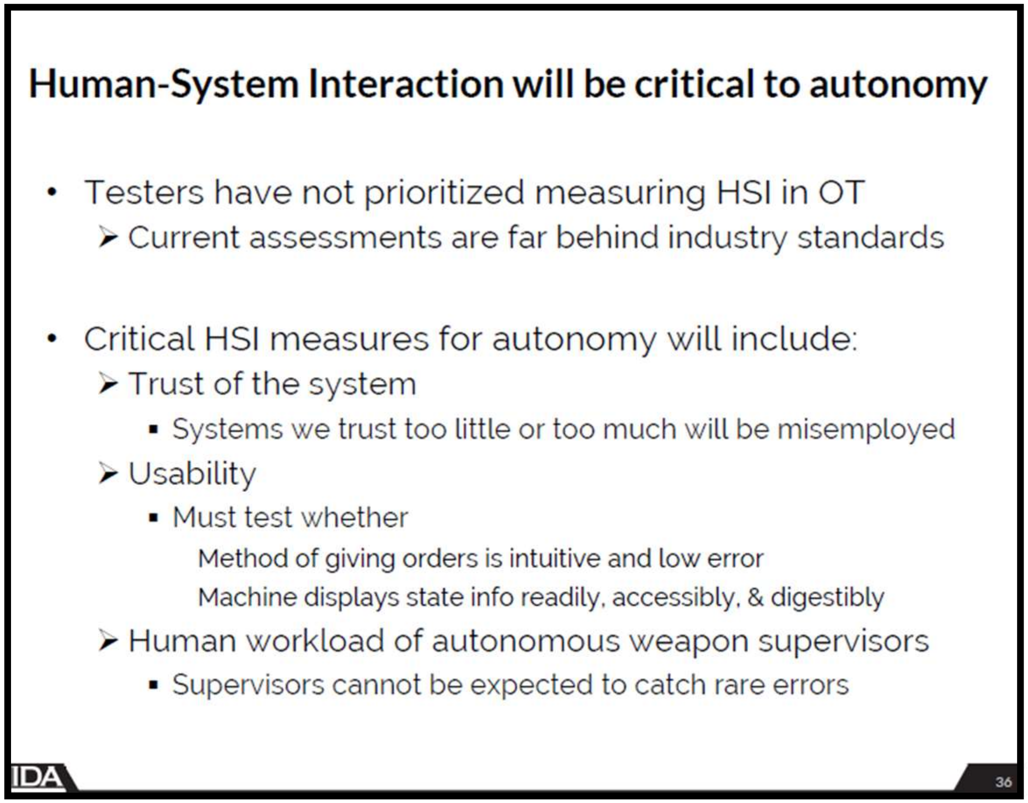
INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-9266

Operational Testing of Systems with Autonomy

Heather M. Wojton, Project Leader

Daniel J. Porter
Yevgeniya K. Pinelis
Chad M. Bieber
Heather M. Wojton
Michael O. McAnally
Laura J. Freeman

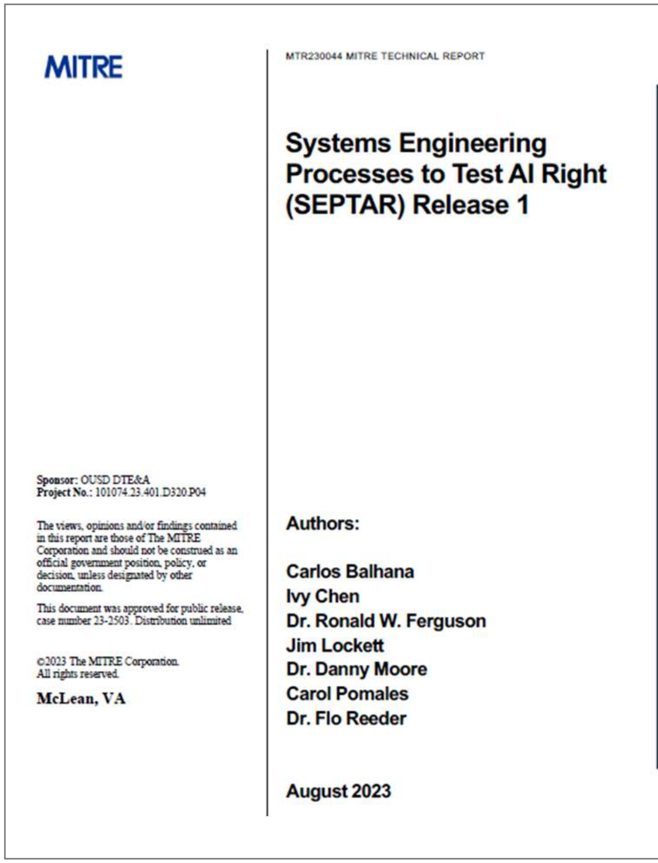


Human-System Interaction will be critical to autonomy

- Testers have not prioritized measuring HSI in OT
  - ➢ Current assessments are far behind industry standards
- Critical HSI measures for autonomy will include:
  - ➢ Trust of the system
    - ▪ Systems we trust too little or too much will be misemployed
  - ➢ Usability
    - ▪ Must test whether
      - Method of giving orders is intuitive and low error
      - Machine displays state info readily, accessibly, & digestibly
  - ➢ Human workload of autonomous weapon supervisors
    - ▪ Supervisors cannot be expected to catch rare errors

36

HUMAN SYSTEMS INTEGRATION FOR DODD 3000.09: MITRE SYSTEMS ENGINEERING



MITRE

MTR230044 MITRE TECHNICAL REPORT

Systems Engineering Processes to Test AI Right (SEPTAR) Release 1

Sponsor: OUSD DTE&A
Project No.: 101074.23.401.D320.P04

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

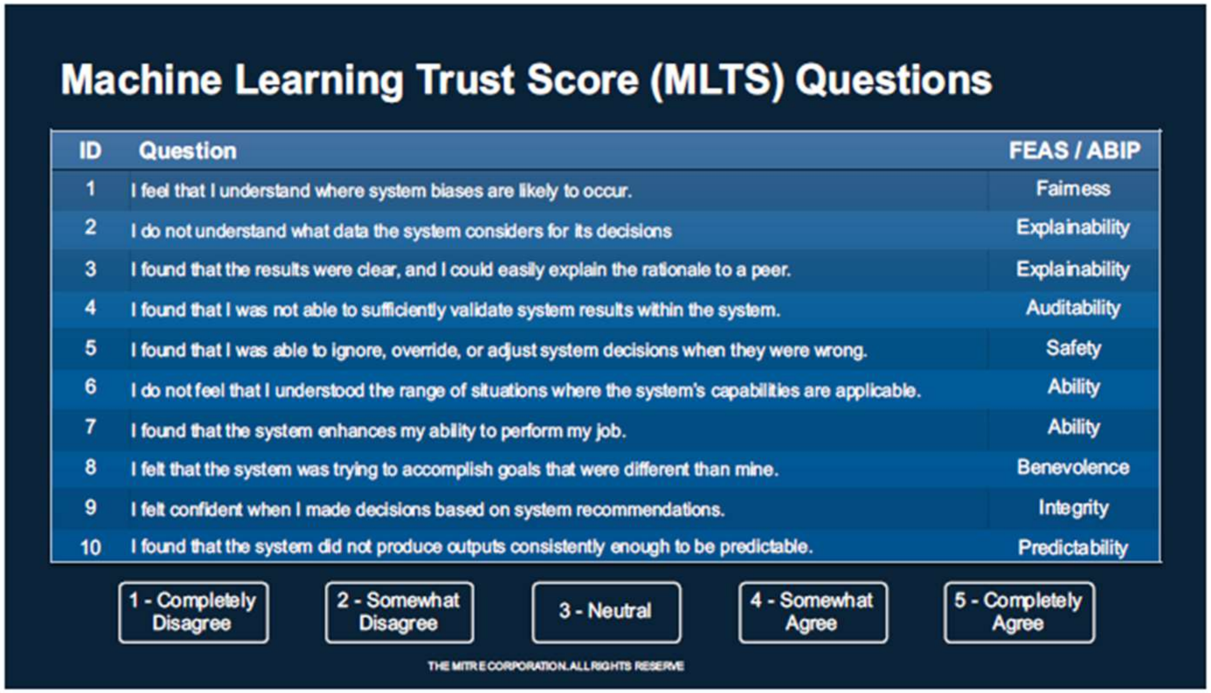
This document was approved for public release, case number 23-2503. Distribution unlimited

McLean, VA

Authors:

Carlos Balhana
Ivy Chen
Dr. Ronald W. Ferguson
Jim Lockett
Dr. Danny Moore
Carol Pomales
Dr. Flo Reeder

August 2023



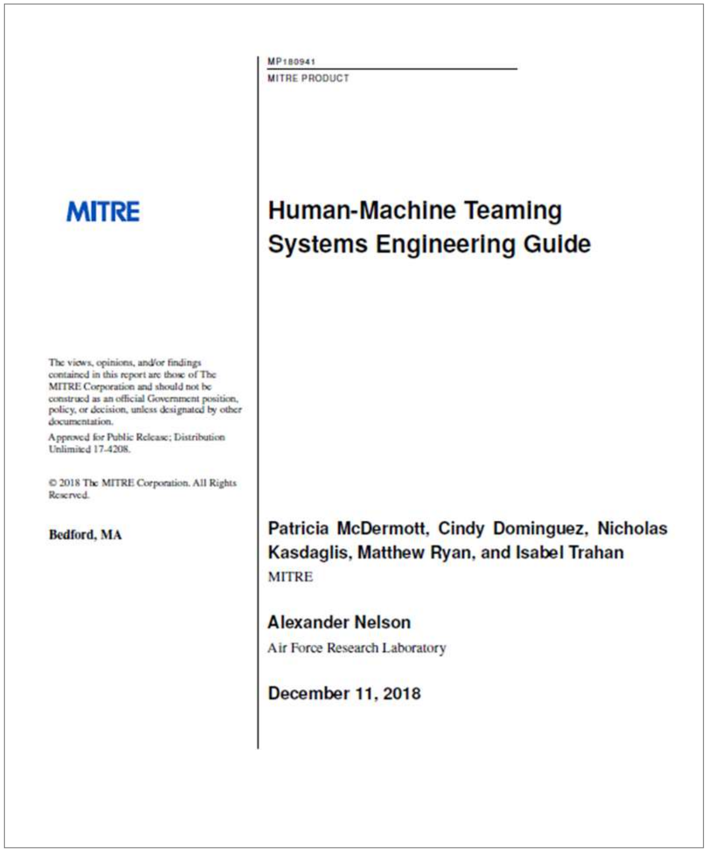
Machine Learning Trust Score (MLTS) Questions

ID	Question	FEAS / ABIP
1	I feel that I understand where system biases are likely to occur.	Fairness
2	I do not understand what data the system considers for its decisions	Explainability
3	I found that the results were clear, and I could easily explain the rationale to a peer.	Explainability
4	I found that I was not able to sufficiently validate system results within the system.	Auditability
5	I found that I was able to ignore, override, or adjust system decisions when they were wrong.	Safety
6	I do not feel that I understood the range of situations where the system's capabilities are applicable.	Ability
7	I found that the system enhances my ability to perform my job.	Ability
8	I felt that the system was trying to accomplish goals that were different than mine.	Benevolence
9	I felt confident when I made decisions based on system recommendations.	Integrity
10	I found that the system did not produce outputs consistently enough to be predictable.	Predictability

1 - Completely Disagree | 2 - Somewhat Disagree | 3 - Neutral | 4 - Somewhat Agree | 5 - Completely Agree

Figure K-1. Machine Learning Trust Score (MLTS) Questions [64]

HUMAN SYSTEMS INTEGRATION FOR DODD 3000.09: MITRE HUMAN MACHINE TEAMING



MP180941

MITRE PRODUCT

MITRE

Human-Machine Teaming Systems Engineering Guide

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

Approved for Public Release; Distribution Unlimited 17-4208.

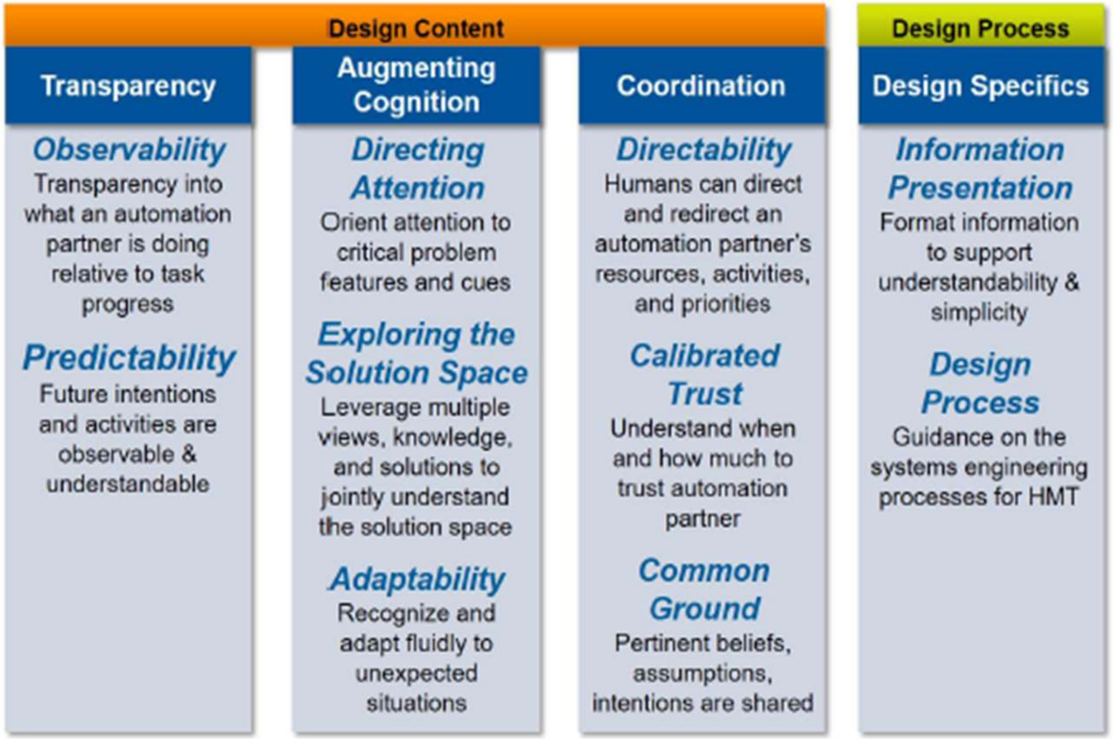
© 2018 The MITRE Corporation. All Rights Reserved.

Bedford, MA

Patricia McDermott, Cindy Dominguez, Nicholas Kasdaglis, Matthew Ryan, and Isabel Trahan
MITRE

Alexander Nelson
Air Force Research Laboratory

December 11, 2018



Design Content

Transparency

Observability Transparency into what an automation partner is doing relative to task progress

Predictability Future intentions and activities are observable & understandable

Augmenting Cognition

Directing Attention Orient attention to critical problem features and cues

Exploring the Solution Space Leverage multiple views, knowledge, and solutions to jointly understand the solution space

Adaptability Recognize and adapt fluidly to unexpected situations

Coordination

Directability Humans can direct and redirect an automation partner's resources, activities, and priorities

Calibrated Trust Understand when and how much to trust automation partner

Common Ground Pertinent beliefs, assumptions, intentions are shared

Design Process

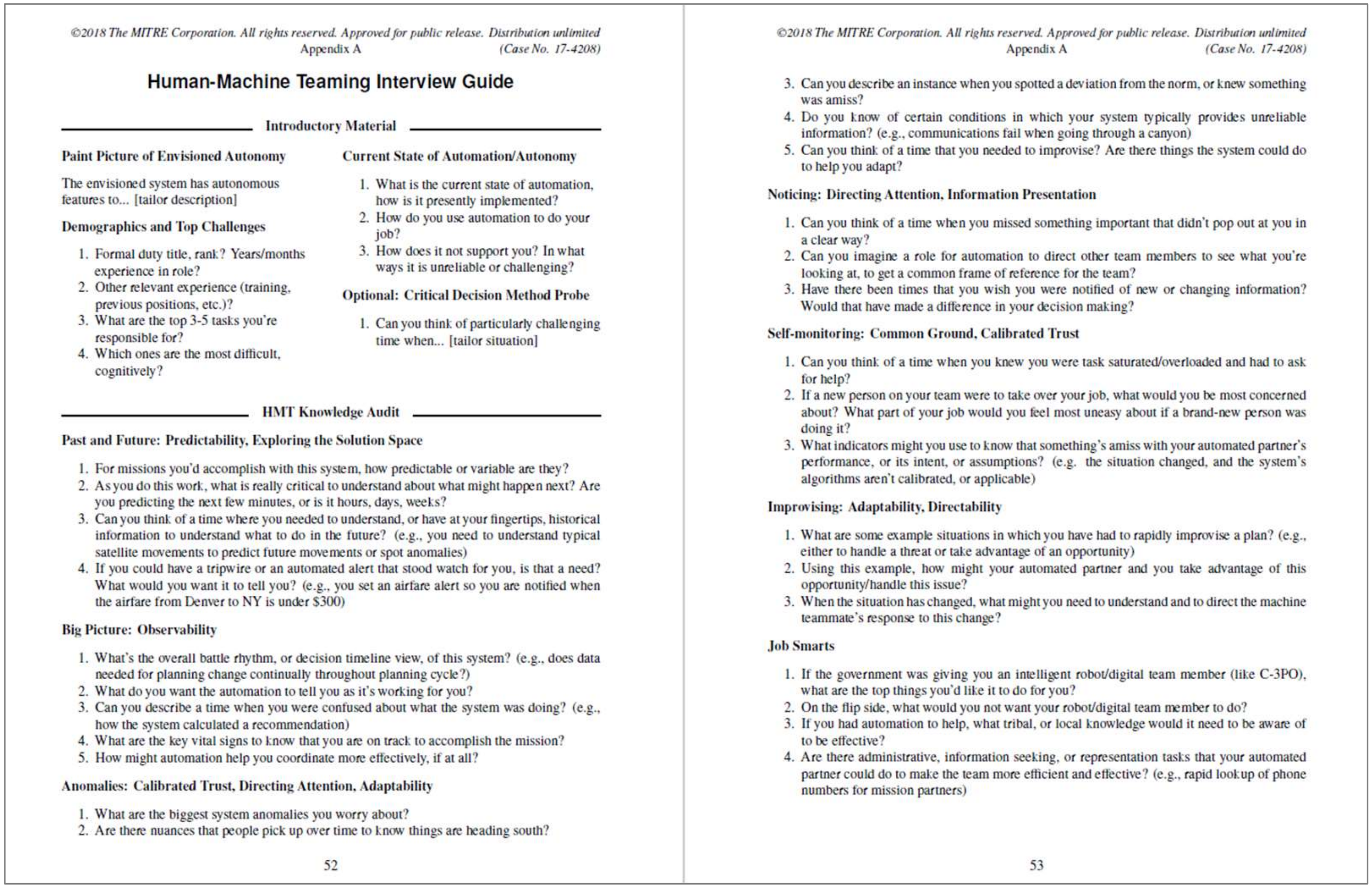
Design Specifics

Information Presentation Format information to support understandability & simplicity

Design Process Guidance on the systems engineering processes for HMT

Figure 2. Ten leverage points organized into a Framework for HMT

HUMAN SYSTEMS INTEGRATION FOR DODD 3000.09: MITRE HUMAN MACHINE TEAMING – CONT'D



Human-Machine Teaming Interview Guide

Introductory Material

Paint Picture of Envisioned Autonomy

The envisioned system has autonomous features to... [tailor description]

Demographics and Top Challenges

1. Formal duty title, rank? Years/months experience in role?
2. Other relevant experience (training, previous positions, etc.)?
3. What are the top 3-5 tasks you're responsible for?
4. Which ones are the most difficult, cognitively?

Current State of Automation/Autonomy

1. What is the current state of automation, how is it presently implemented?
2. How do you use automation to do your job?
3. How does it not support you? In what ways it is unreliable or challenging?

Optional: Critical Decision Method Probe

1. Can you think of particularly challenging time when... [tailor situation]

HMT Knowledge Audit

Past and Future: Predictability, Exploring the Solution Space

1. For missions you'd accomplish with this system, how predictable or variable are they?
2. As you do this work, what is really critical to understand about what might happen next? Are you predicting the next few minutes, or is it hours, days, weeks?
3. Can you think of a time where you needed to understand, or have at your fingertips, historical information to understand what to do in the future? (e.g., you need to understand typical satellite movements to predict future movements or spot anomalies)
4. If you could have a tripwire or an automated alert that stood watch for you, is that a need? What would you want it to tell you? (e.g., you set an airfare alert so you are notified when the airfare from Denver to NY is under $300)

Big Picture: Observability

1. What's the overall battle rhythm, or decision timeline view, of this system? (e.g., does data needed for planning change continually throughout planning cycle?)
2. What do you want the automation to tell you as it's working for you?
3. Can you describe a time when you were confused about what the system was doing? (e.g., how the system calculated a recommendation)
4. What are the key vital signs to know that you are on track to accomplish the mission?
5. How might automation help you coordinate more effectively, if at all?

Anomalies: Calibrated Trust, Directing Attention, Adaptability

1. What are the biggest system anomalies you worry about?
2. Are there nuances that people pick up over time to know things are heading south?

3. Can you describe an instance when you spotted a deviation from the norm, or knew something was amiss?
4. Do you know of certain conditions in which your system typically provides unreliable information? (e.g., communications fail when going through a canyon)
5. Can you think of a time that you needed to improvise? Are there things the system could do to help you adapt?

Noticing: Directing Attention, Information Presentation

1. Can you think of a time when you missed something important that didn't pop out at you in a clear way?
2. Can you imagine a role for automation to direct other team members to see what you're looking at, to get a common frame of reference for the team?
3. Have there been times that you wish you were notified of new or changing information? Would that have made a difference in your decision making?

Self-monitoring: Common Ground, Calibrated Trust

1. Can you think of a time when you knew you were task saturated/overloaded and had to ask for help?
2. If a new person on your team were to take over your job, what would you be most concerned about? What part of your job would you feel most uneasy about if a brand-new person was doing it?
3. What indicators might you use to know that something's amiss with your automated partner's performance, or its intent, or assumptions? (e.g. the situation changed, and the system's algorithms aren't calibrated, or applicable)

Improvising: Adaptability, Directability

1. What are some example situations in which you have had to rapidly improvise a plan? (e.g., either to handle a threat or take advantage of an opportunity)
2. Using this example, how might your automated partner and you take advantage of this opportunity/handle this issue?
3. When the situation has changed, what might you need to understand and to direct the machine teammate's response to this change?

Job Smarts

1. If the government was giving you an intelligent robot/digital team member (like C-3PO), what are the top things you'd like it to do for you?
2. On the flip side, what would you not want your robot/digital team member to do?
3. If you had automation to help, what tribal, or local knowledge would it need to be aware of to be effective?
4. Are there administrative, information seeking, or representation tasks that your automated partner could do to make the team more efficient and effective? (e.g., rapid lookup of phone numbers for mission partners)

OTHER BEHAVIORAL TESTS AND COMPONENTS

- Trust
  - Validated measures of trust
    - Trust in Automation Scale (TAS) (Jian et al., 2000)
    - Trust of Automated Systems Test (TOAST) (Wojton et al., 2020)
    - Trust Perception Scale-HRI (HRI) (Schaefer, 2016)
- Workload
  - NASA Task Load Index (NASA- TLX) (Hart, 2006)
  - Psychophysiological Measures
- Useability
  - System Usability Scale (SUS) (Brooke, 1996)
- Soldier Acceptance
  - Technology Acceptance Model (TAM) (Davis, 1989)

REPOSITORIES OF TESTS

- The Institute for Defense Analyses maintains an archive of Department of Operational Testing and Evaluation recommended human system integration scales at the DOT&E Validated Scale Repository (https://testscience.org/validated-scales-repository-intro/). Scales that assess usability, workload, and user trust are available, including the questionnaire items, administration instructions, scoring, and listed advantages and disadvantages.
- The Joint Human Systems Integration Working Group (JHSIWG) under the auspices of the Office of the Under Secretary of Defense-Research and Engineering, Specialty Engineering maintains a searchable database of human systems engineering tools used that may be relevant to testing and evaluation of autonomous weapon systems. Categories of human systems engineering tools include those for human factors and ergonomics, situational awareness, and workload. The archive is housed in on the APAN system and requires registration (https://sites.apan.org/osd/HSI-BOKM/default.aspx).
- The Chief Digital and Artificial Intelligence Office maintains an online tailorable form to guide developers of AI with tools, assessments, and artifacts. The Responsible Artificial Intelligence (RAI) Toolkit contains 106 tools to assist in mitigating risks or improving development of AI systems (https://rai.tradewindai.com/tools-list). Tools listed on the site for human systems integration include “HMT Guidebook”, “RAI UX/HMT Toolkit”, the Human-Machine Teaming Systems Engineering Guide”, “Trust in Autonomous Systems Test”, “System Usability Scale”, and “Human AI control research instrument”. The site continues to evolve; therefore, additional relevant to assessing appropriate levels of human judgement may appear as well.

DODD 3000.09 HUMAN SYSTEMS INTEGRATION CATEGORIES

- Link between Operator and System: Displays
  - System Status
  - Target Status
  - Artificial Intelligence Confidence (confidence of AI in decision)
  - Collateral Situation
  - Situational Awareness
- Link between Operator and System: Inputs/Outputs
  - Form Factor
  - Physical Layout
  - Effectors (e.g., buttons, joystick)
  - Screen Layout
  - Menu Configuration
  - Psychomotor Limitations
  - Task Guidance
- Warfighter Responses
  - Trust
  - Reliance
  - Confidence
  - Acceptance
  - Complacency
  - Vigilance
  - Attention
  - Fatigue
  - Stress
- Cognitive Alignment
  - Mental Models
  - Common Knowledge
  - Transparency
  - Explainability

T&E FOR LINK BETWEEN OPERATOR AND SYSTEM: DISPLAYS PART 1

LINK BETWEEN OPERATOR AND SYSTEM: DISPLAY PART 1	
SYSTEM STATUS	
Utility of Information	Interview, Ratings after simulation
Interpretability/Clarity of Presentation	Performance in simulation, Ratings
Out of Boundary Condition Warnings	Frequency of detection of out of boundary conditions.
TARGET STATUS (FIND, FIX, TRACK, TARGET, ENGAGE, ASSESS)	
Utility of Information	Performance in simulation, Ratings, Interview
Interpretability/Clarity of Presentation	Performance in simulation, Ratings, Interview
ARTIFICIAL INTELLIGENCE CONFIDENCE	
Utility of Information	Interview, Ratings
Interpretability/Clarity of Presentation	Interview, Ratings

T&E FOR LINK BETWEEN OPERATOR AND SYSTEM: DISPLAYS PART 2

LINK BETWEEN OPERATOR AND SYSTEM: DISPLAY PART 2	
COLLATERAL SITUATION	
Utility of Information	Performance in simulation (acceptability of collateral damage), Ratings, Interview
Interpretability/Clarity of Presentation	Performance in simulation (acceptability of collateral damage), Ratings, Interview
SITUATIONAL AWARENESS	
Utility of Information	Performance in simulation, Ratings, Interview
Interpretability/Clarity of Presentation	Performance in simulation, Ratings, Interview

T&E FOR LINK BETWEEN OPERATOR AND SYSTEM: INPUTS/OUTPUTS PART 1

LINK BETWEEN OPERATOR AND SYSTEM: INPUTS/OUTPUTS PART 1	
Form Factor	Interview, ratings, performance time during simulation testing
Physical Layout	Interview, ratings, performance time during simulation testing, ergonomic testing, SUS, TAM
Effectors (e.g., buttons, joystick, etc.)	Interview, ratings, performance time during simulation testing, ergonomic testing, SUS, TAM
Screen Layout	Interview, ratings, performance time during simulation testing, SUS, TAM
Menu Configuration	Interview, ratings, performance time during simulation testing, SUS, TAM
Psychomotor Limitations	Performance time during simulation testing

T&E FOR LINK BETWEEN OPERATOR AND SYSTEM: INPUTS/OUTPUTS PART 2

LINK BETWEEN OPERATOR AND SYSTEM: INPUTS/OUTPUTS PART 2	
TASK GUIDANCE	
What Operator Does	Interview, ratings, performance time during simulation testing
What System Does	Interview, ratings, performance time during simulation testing
Activation	Time to activation in simulation testing
Deactivation	Time to deactivation in simulation testing
Emergency Stops	Time from warning to stop in simulation testing

T&E FOR WARFIGHTER RESPONSE PART 1

WARFIGHTER RESPONSE PART 1	
WORKLOAD	
Physical	Interview, NASA TLX, psychophysiological monitoring, performance in simulation testing
Cognitive	Interview, NASA TLX, psychophysiological monitoring, performance in simulation testing, MCH
Temporal	Interview, NASA TLX, timing during simulation testing
USABILITY	
Utility	SUS, interview, subject matter expert evaluation, performance in simulation testing
Ease of Use	Percentage of SUS, interview, subject matter expert evaluation, performance in simulation testing.

T&E FOR WARFIGHTER RESPONSE PART 2

WARFIGHTER RESPONSE PART 2	
Trust	TOAST, HRI, TAS
Reliance	Percentage of time operator choose to use system rather than alternative in simulation testing.
Confidence	Survey
Acceptance	TAM
Complacency	Behavioral observation (choice to act)
Vigilance	Behavioral observation (time watching on screen), eye tracking.
Attention	Behavioral observation, detection of events of note, eye tracking.
Fatigue	Survey, NASA-TLX, psychophysiological measures
Stress	Survey, NASA-TLX, psychophysiological measures

T&E FOR COGNITIVE ALIGNMENT

COGNITIVE ALIGNMENT	
Mental Models	Survey, interview, errors occurring during simulation testing that might point to mismatch in mental models.
Common Knowledge	Survey, interview, errors occurring during simulation testing that might point to mismatch in mental models.
Transparency	Survey, interview, errors occurring during simulation testing that might point to lack of knowledge of AI operations or processes.
Explainability	Survey, interview, errors occurring during simulation testing that might point to lack of knowledge of the rationale behind a decision made by the AI.

SUMMARY

- Testing and evaluation for “appropriate levels of human judgement” is essentially assurance of optimal human systems integration.
- CDAO, IDA, MITRE have proposed testing general frameworks for testing and evaluation for a broad set of artificial intelligence and autonomous systems.
- DoDD 3000.09 identifies specific information requirements for human systems integration that can be addressed with existing tools.
- Implementation of DoDD 3000.09 systems requires incorporation of human systems integration/human factors scientists and engineers.
- Questions?

THANK YOU.

Elizabeth Mezzacappa, PhD
Tactical Behavior Research Laboratory
US Army Combat Capabilities Development Command Armaments Center
Picatinny Arsenal, NJ USA
elizabeth.s.mezzacappa.civ@army.mil
elizabeth.s.mezzacappa.civ@mail.smil.mil
(520) 684-2830

U.S. ARMY