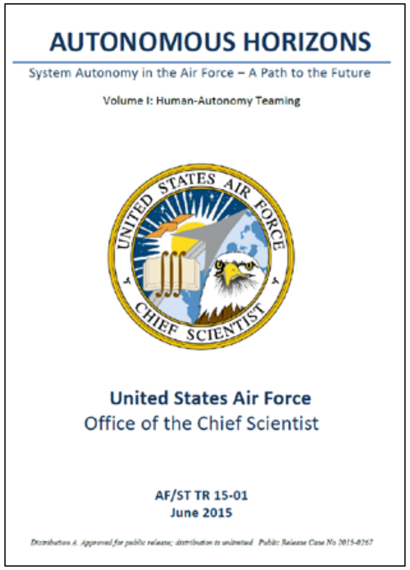
Director,
Operational Test and Evaluation

Human-Autonomy Teaming:
T&E Issues and Recommendations

Dr. Greg Zacharias
Chief Scientist
Operational Test and Evaluation
Office of Secretary of Defense

28 JUL 2021

Towards a Three-Volume Set



AUTONOMOUS HORIZONS

System Autonomy in the Air Force – A Path to the Future

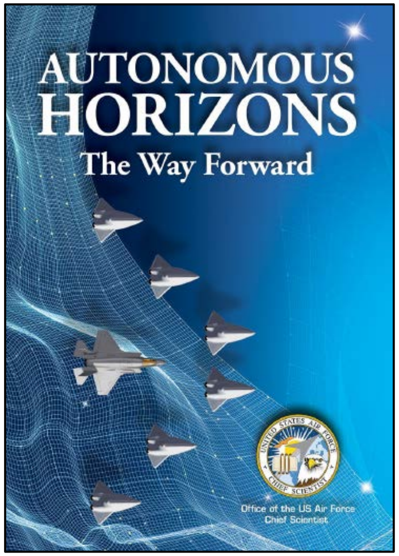
Volume I: Human-Autonomy Teaming

United States Air Force
Office of the Chief Scientist

AF/ST TR 15-01
June 2015

Distribution A. Approved for public release; distribution is unlimited. Public Release Case No 2015-0267

A vision for autonomous systems working synergistically with our airmen, enabling human-autonomy teaming with seamless situation awareness, decisions, and actions

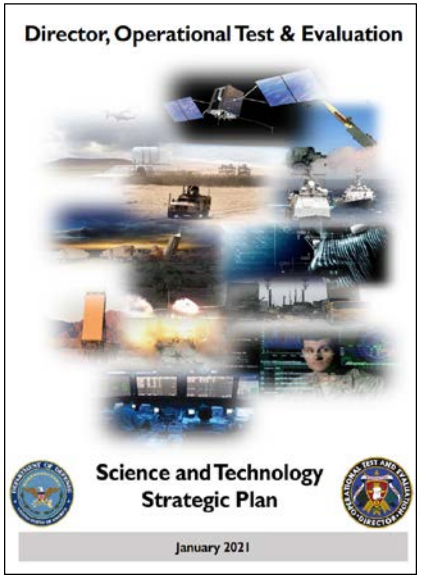


AUTONOMOUS HORIZONS

The Way Forward

Office of the US Air Force Chief Scientist

An overview of the technical issues in creating machine intelligence to deal with the challenges of uncertainty & variability in operational environments



Director, Operational Test & Evaluation

Science and Technology Strategic Plan

January 2021

Key development issues including cyber security, command & control, counterautonomy, and test and evaluation (T&E)

DOT&E Activities and Mission

-Advise on testable, mission-relevant requirements

-Approve Test & Evaluation Master Plan submitted by Program Office

-Approve operational and live fire Test Plans submitted by Service OTAs

-Collaborate with DT&E to gain early insight into performance

-Evaluate system performance in a report to Congress & DoD leadership

-Inform production/fielding decisions

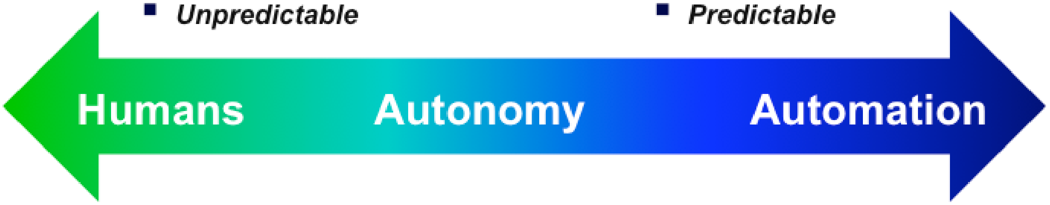
The short version…

User-Centered Design → User-Centered Test and Evaluation

OTA: Operational Test Agency
DT&E: Developmental Test and Evaluation

Key Human-Autonomy Teaming Issues (Volume 1)

What’s the relationship between humans, autonomy, and automation?



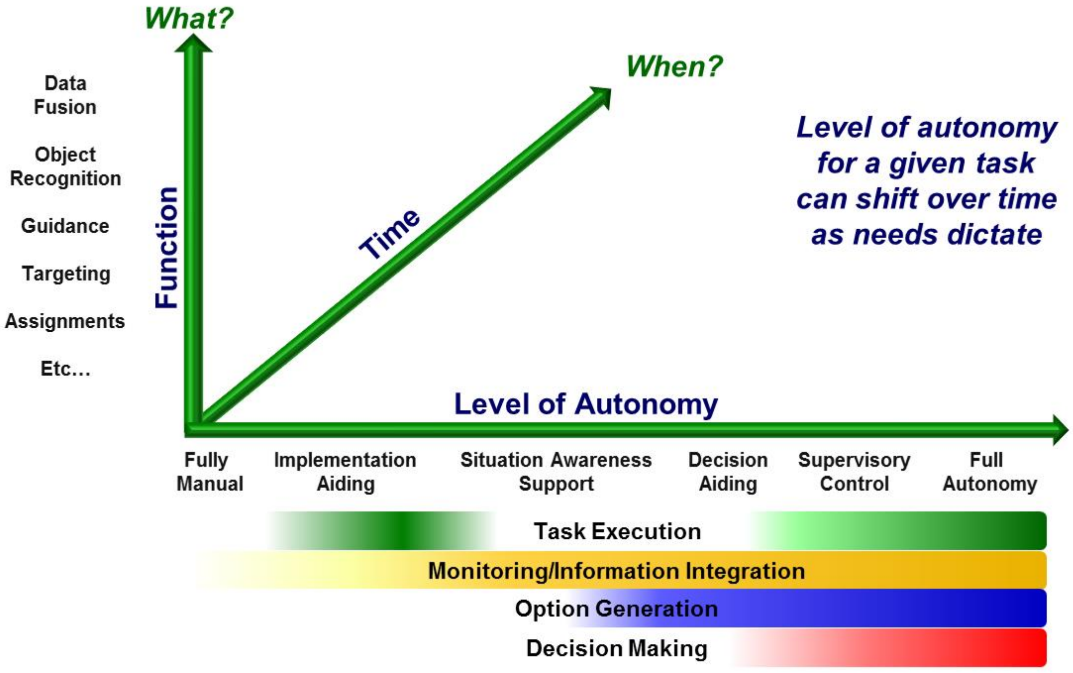
- *Novel situations*
- *Moderate repeatability*
- *May be inconsistent*
- *Unpredictable*

- *Constrained situations*
- *Reliable*
- *Consistent*
- *Predictable*

Humans

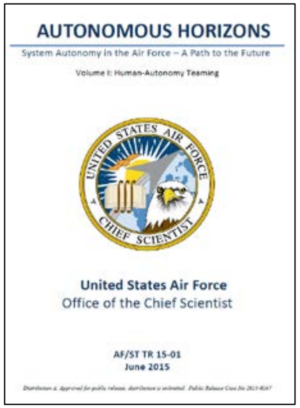
Autonomy

Automation



Level of autonomy for a given task can shift over time as needs dictate

What levels of autonomy apply to which human-system teaming functions, and how can these change over time?



AUTONOMOUS HORIZONS

System Autonomy in the Air Force – A Path to the Future

Volume I: Human-Autonomy Teaming

United States Air Force
Office of the Chief Scientist

AF/ST TR 15-01
June 2015

Distribution A. Approved for public release; distribution is unlimited. Public Release Case No [illegible]

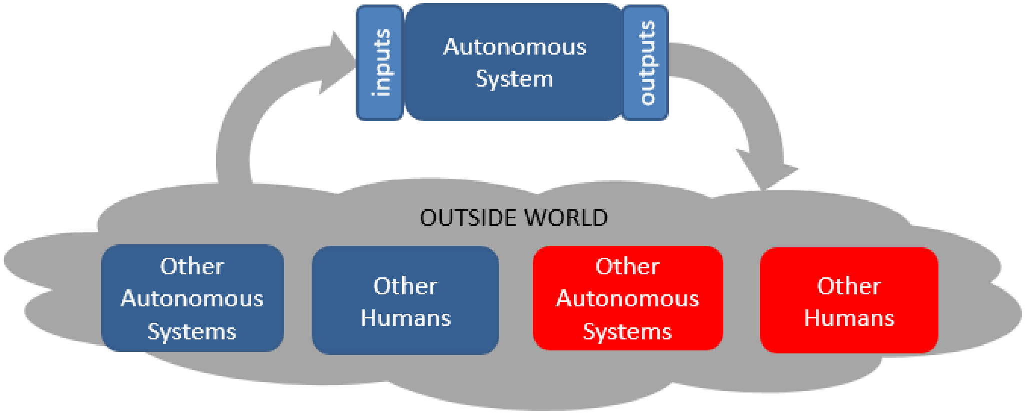
Key Autonomous Systems Attributes (Volume 2)

Situated Agency

- Sensing the environment, assessing the situation, reasoning about it, making decisions to reach a goal, and then acting on it

Adaptive Cognition

- Using different modes of "thinking", from low-level rules, to high-level reasoning



Multi-Agent Emergence

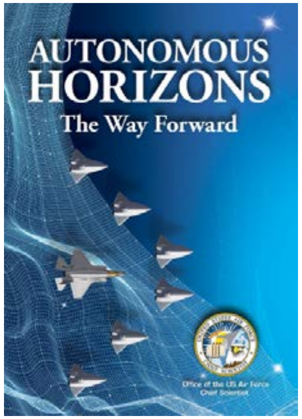
- Interacting with other agents, human or otherwise, affording novel emergent behavior of the group/team

Experiential Learning

- "Learning" new behaviors over time and experience…

Desired properties

- Proficiency, trustworthiness, flexibility → AI-Enabled



T&E Concerns: Some Studies

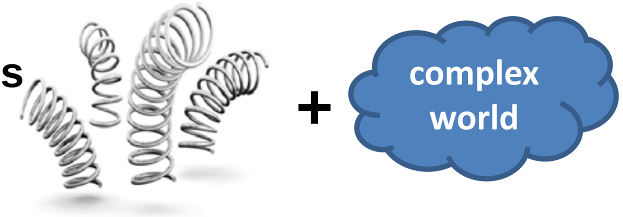
- T. Menzies and C Pecheur, *Verification and Validation and Artificial Intelligence*, Preprint submitted to Elsevier Science, 12 July 2004
- W. Dahm, 2011. *Technology Horizons: A Vision for Air Force Science and Technology 2010-30*. Maxwell AFB, AL: Air University Press.
- V. Roske, I. Kohlberg, and R. Wagner, *Autonomous Systems Challenges to Test and Evaluation*, National Defense Industrial Association, Test and Evaluation Conference, 12-15 March 2012
- DSB, 2012 Defense Science Board Autonomy Study: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, *The Role of Autonomy in DoD Systems,* Washington, DC, 2012
- DOD, DOD R&E Autonomy Community of Interest, T&E V&V (TEVV) Working Group, *Technology Investment Strategy 2015-2018,* OASD(R&E), May 2015
- D. Ahner and C. Parson, *Workshop Report: Test and Evaluation of Autonomous Systems*, STAT Center of Excellence, Wright-Patterson AFB, OH, 2016
- DSB, 2016 Defense Science Board Autonomy Study: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, *Report of the Defense Science Board Summer Study on Autonomy,* Washington DC, 2016
- A. Hill and G. Thompson, *FIVE GIANT LEAPS FOR ROBOTKIND: EXPANDING THE POSSIBLE IN AUTONOMOUS WEAPONS*, War on the Rocks, https: ns/, 28 DEC 2016
- SAB, US Air TR-17-03, 15 September 2
- B. Haugh, D. *Systems*, P-9292, Institute for Defense Analysis, Alexandria, VA, 2018
- T. Talafuse and D. Ahner, *Workshop Report: Test and Evaluation of Autonomous Systems,* Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE), Air Force Institute of Technology, March 2018
- D. K. Ahner, C. R. Parson, J. L. Thompson, and W. F. Rowell, *Overcoming the Challenges in Test and Evaluation of Autonomous Robotic Systems*, ITEA J. of Test and Evaluation, 39: 86-94, June 2018
- A. L. McLean, J. R. Bertram, J. A. Hoke, S. S. Rediger, and J. C. Skarphol, *LVC-Enabled Testbed for Autonomous System Testing*, ITEA J. of Test and Evaluation, 39: 120-128, June 2018
- P. Caseley, *Human-Machine Trust: Risk-Based Assurance and Licensing of Autonomous Systems, SCI-313 Specialist Meeting Report*, NATO STO-MP-SCI-313, 3-5 December 2018
- H. Miller, *Senate Report on Test Infrastructure: Autonomy*, MITRE, April 2019 [need updated version]
- J.C. Lede, *Autonomy Overview, US-Japan Service to Service Dialog*, Autonomy Community of Interest Lead, AFRL, OSD, 5 APR 2019
- Y. Gil and B. Selman, *A 20-Year Community Roadmap for Artificial Intelligence Research in the US*, American Association for Artificial Intelligence Draft Report, May 2019

And more are on the way (!)

AS Test and Evaluation: Issues (1 of 3)

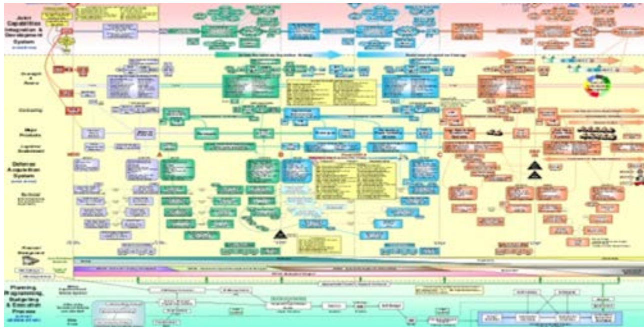
“Flexible” ASs operating in complex, stochastic, dynamic environments

- Flexibility+complexity → ill-defined requirements / → huge state-spaces
- External variability + internal complexities
- Learning and emergence



Acquisition pipeline unready for these systems

- Requirements needed at operational/behavioral level… with traceability through CT/DT/OT
- Rigid processes for evolving systems
- Few common T&E processes and data formats



Infrastructure shortcomings hamper AS development *and* T&E

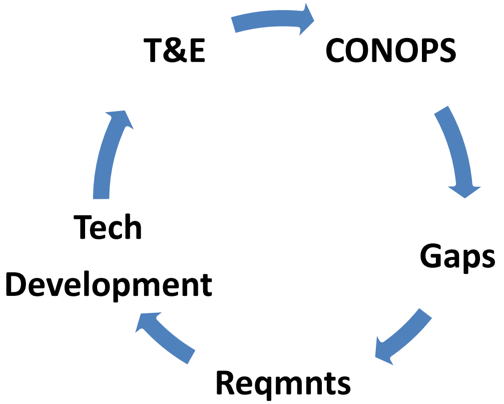
- Lack of common AS frameworks/architectures
- Little/no instrumentation or “design for testability”
- Current manual certification methods limited
- Lack of T&E testbeds, ranges, personnel

CT/DT/OT: Contractor Test, Developmental Test, Operational Test

AS Test and Evaluation: Issues (2 of 3)

Conflation of technology and CONOPS

- S&T traditionally driven by existing CONOPS
- But AS proficiency/flexibility will drive new CONOPS
- Expect conflation of CONOPS, AS development, *and* T&E
- Compounded by prototyping of systems and CONOPS experimentation



Inadequacy of traditional T&E methods & tools

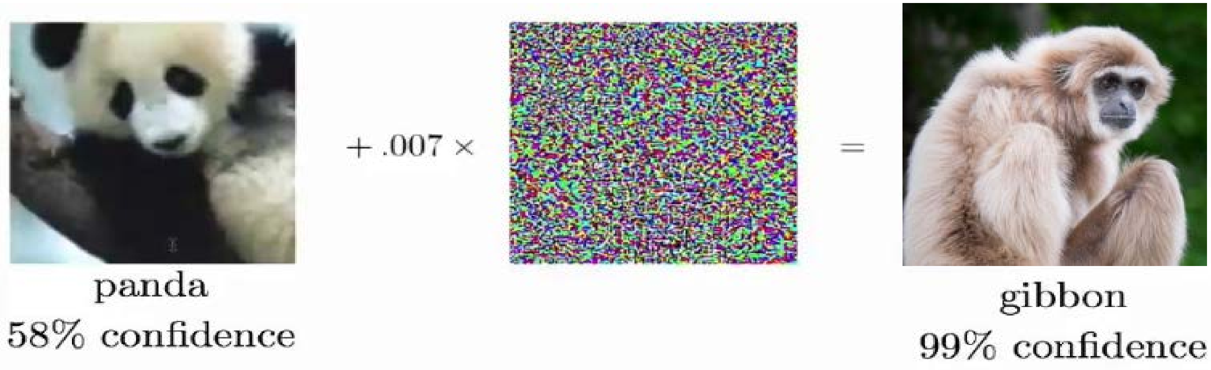
- ASs likely to *learn* with training, experience, and cultural (fleet) learning
- But current T&E methods don't deal well with changing systems under test (SUTs)
- ASs likely to interact with their AS peers, leading to *emergent* behaviors
- Learning *and* emergence → hard for human evaluator (*and* teammate) to track what AS is doing, let alone design/conduct effective and rigorous T&E

S&T: Science and Technology; CONOPS: Concept of Operations

AS Test and Evaluation: Issues (3 of 3)

Unique T&E challenges to ensuring safe and secure operations

- Conventional cyber attacks “tuned” for subtle effects on perception, decision-making, …
- Adversarial AI attacks can degrade performance, cause errors, or trigger unwanted behaviors
- Like pre-developmental “data poisoning”
- Or post-deployment real-time counter autonomy attacks (Goodfellow, 2016)

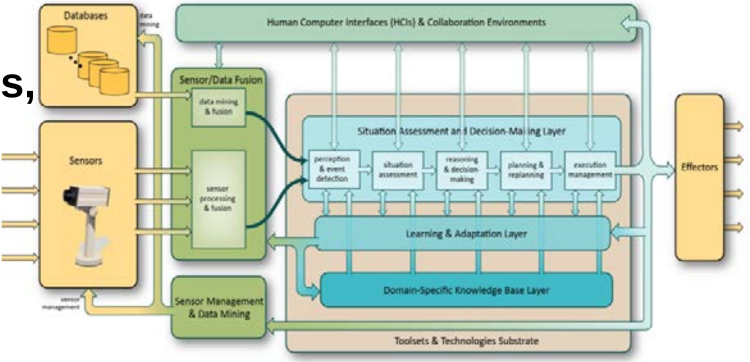


Real-time monitoring systems for safe operations bring their own T&E demands → “homunculi all the way up”

AS Test and Evaluation: Recommendations (1 of 2)

Requirements, design, and development

- Architect ASs using common frameworks and modular subsystems
- Support “cognitive instrumentation” via sensors, assessors, and “explainers”
- Follow accepted HSI design principles
- Curate the data used for training; protect from “poisoning”; enrich for robust response
- Invest in modeling and simulation-based T&E


Databases
data mining
Human Computer Interfaces (HCIs) & Collaboration Environments
Sensor/Data Fusion
data mining & fusion
sensor processing & fusion
Sensors
sensor management
Situation Assessment and Decision-Making Layer
perception & event detection
situation assessment
reasoning & decision making
planning & replanning
execution management
Effectors
Learning & Adaptation Layer
Domain-Specific Knowledge Base Layer
Sensor Management & Data Mining
Toolsets & Technologies Substrate

Extend existing and develop new T&E methods/ tools to deal with complex/stochastic/emergent behaviors, and AS-specific vulnerabilities

- Research/embrace new methods/tools for complex, stochastic, and non-stationarity systems
- Develop new statistical engineering methods for T&E design and analysis
- Extend nascent efforts in human-machine interaction and human-AS teaming
- Account for “emergent behavior” across systems and the impact on the SUT
- Assess cyber vulnerabilities and adversarial attack effects/mitigators

HSI: Human-System Integration; SUT: System Under Test

AS Test and Evaluation: Recommendations (2 of 2)

Infrastructure and process

- Move to a “T&E lifecycle” viewpoint/culture
  - Break stovepipes and reduce CT/DT/OT cycles while preserving legal firewalls
- Invest in “digital modernization”
  - Develop unifying infrastructure for requirements generation/traceability
  - Integrate heterogeneous test data via common data formats and networks
- Make massive use of M&S, test automation, & data analytics everywhere

Risk assurance

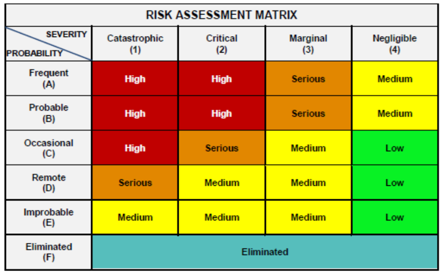
- AS training: curate, protect, “robustify” data
- Augment subjective risk assessments with formal assurance arguments
- Shape requirements setting with risk assessments

Human-autonomy teaming

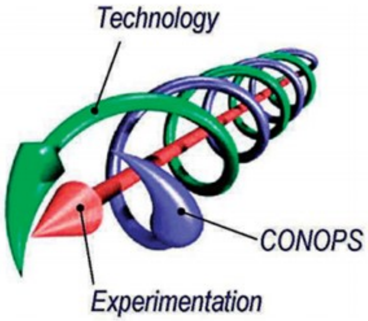
- Embrace co-development of CONOPS with ASs
- Measure adherence to HSI design principles
- Emphasize pre-test training/teaming

DoD DIGITAL MODERNIZATION STRATEGY

DoD Information Resource Management Strategic Plan FY19-23


RISK ASSESSMENT MATRIX

SEVERITY / PROBABILITY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	High	High	Serious	Medium
Probable (B)	High	High	Serious	Medium
Occasional (C)	High	Serious	Medium	Low
Remote (D)	Serious	Medium	Medium	Low
Improbable (E)	Medium	Medium	Medium	Low
Eliminated (F)	Eliminated			

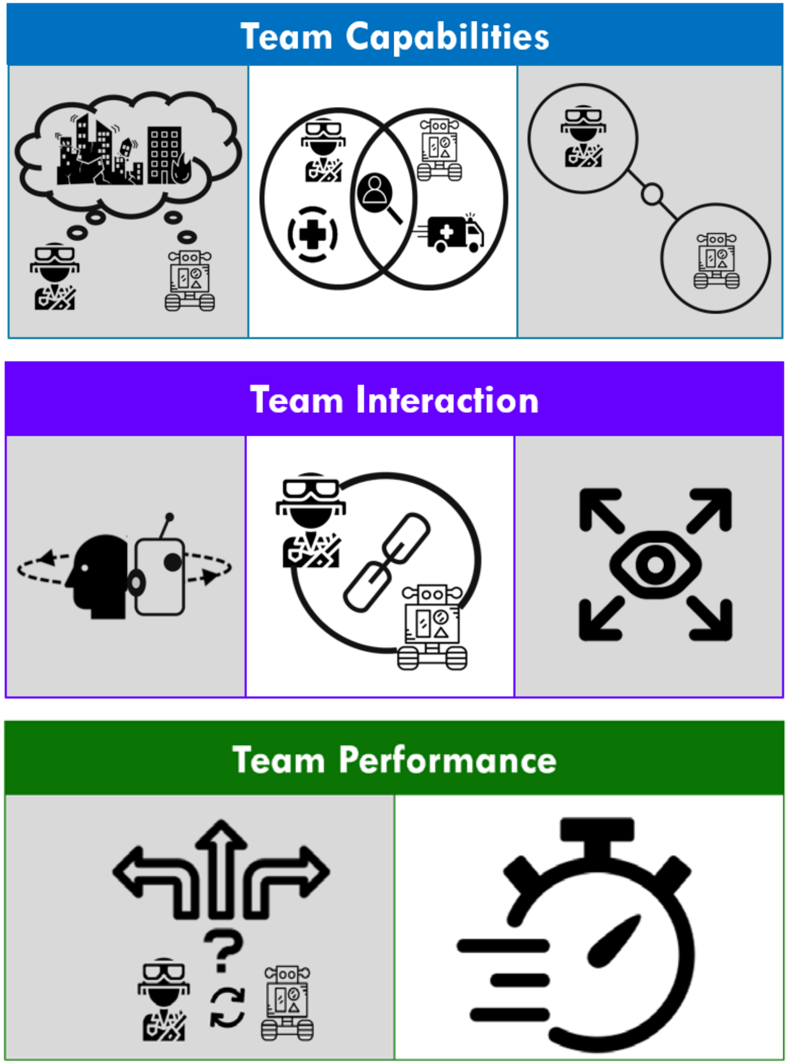


Framework for Human-Autonomy Teaming

This framework:

- Gives specific direction on teaming factors
- Enables tests of whether a team is effective in general, not just during the observed task

Image Credits: 1. *Are Drones Changing the Way We Live?*, D!gitalist, Nov. 2019, https://www.digitalistmag.com/digital-economy/2019/11/05/are-drones-changing-way-we-live-06201367/



Team Capabilities

Team Interaction

Team Performance

One More Thing: Responsible AI (RAI)

Equitable: [DoD] will… minimize unintended bias in AICs [AI capabilities].

Reliable: [DoD's] AICs will have explicit, well-defined uses, and the [associated] safety, security, and effectiveness will be subject to testing and assurance within those defined uses across their entire life-cycles.

Governable: The Department will design and engineer AICs to fulfill their intended functions [and] detect and avoid unintended consequences… [D]eployed systems that demonstrate unintended behavior [will be capable of being disengaged or deactivated.]

Responsible: DoD personnel will …[remain] responsible for the development, deployment, and use of AICs.

Traceable: [DoD's] AICs will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AICs … with transparent and auditable methodologies, data sources, and design procedure and documentation.

"I guess it's ethical. Let me run it through my 'Ethics Check' app."

Next Steps for DOT&E

Short term

- Instances of "partial autonomy" at the component level in test plans are now coming through the office
- Working to develop interim guidelines for dealing with these

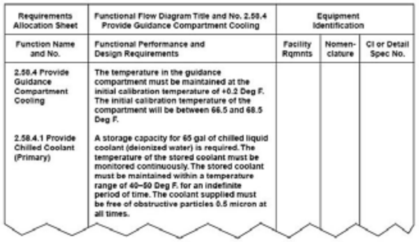
Mid term

- This trend will accelerate
- Working with multiple AI/AS T&E groups throughout DOD covering policy, guidance, technologies, testbeds, and workforce
- Reaching out to all of you in how to deal with this nascent technology
- Need to execute smartly on the recommendations to get ahead of the expected T&E challenges

Backups

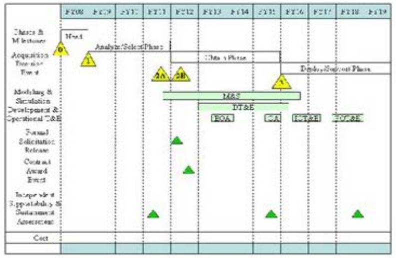
DOT&E Activities and Mission

Advise on Testable, Mission-Relevant Requirements

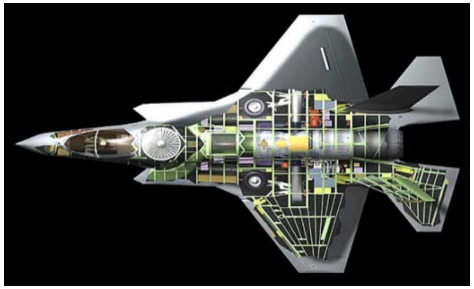


Requirements Allocation Sheet	Functional Flow Diagram Title and No. 2.58.4 Provide Guidance Compartment Cooling	Equipment Identification		
Function Name and No.	Functional Performance and Design Requirements	Facility Rqmnts	Nomen-clature	CI or Detail Spec No.
2.58.4 Provide Guidance Compartment Cooling	The temperature in the guidance compartment must be maintained at the initial calibration temperature of +0.2 Deg F. The initial calibration temperature of the compartment will be between 66.5 and 68.5 Deg F.			
2.58.4.1 Provide Chilled Coolant (Primary)	A storage capacity for 65 gal of chilled liquid coolant (deionized water) is required. The temperature of the stored coolant must be monitored continuously. The stored coolant must be maintained within a temperature range of 40–50 Deg F. for an indefinite period of time. The coolant supplied must be free of obstructive particles 0.5 micron at all times.			

Approve Test & Evaluation Master Plan Submitted by Program Office



Collaborate with DT&E to gain early insight into performance



Approve operational and live fire test plans submitted by Service OTAs

Evaluate system performance in a report to congress & DoD leadership

Inform Production/Fielding Decision

Authoritative source for DoD weapon systems' operational capabilities

Autonomous Systems: T&E Issues

"Flexible" ASs operating in complex, dynamic, stochastic environments

- External variability + internal complexities → huge non-convex state spaces
- Learning over time and experience can change behaviors → non-stationarity
- Emergence of behaviors across agents → potential for changing CONOPS

Infrastructure shortcomings

- Difficulty specifying requirements at an operational/behavioral level
- Acquisition pipeline fundamentally materiel-oriented
- Lack of common AS architectures/frameworks
- Lack of T&E methods, tools, testbeds, ranges, and experienced personnel
- No up-front instrumentation or design for "testability" or "explainability"
- Current certification methods predominantly manual, subjective, specialized

Unique T&E challenges ensuring safety and security

- Real-time monitoring systems for safe operations bring own T&E demands
- Conventional cyber attacks can be "tuned" for subtle attacks on performance
- And adversarial attacks call for expanded T&E scope to better model threats

AS: Autonomous System

Autonomous Systems: T&E Recommendations

T&E needs to influence requirements, design, and development

- Architect ASs using common frameworks and modular subsystems
- Support "cognitive instrumentation" via sensors, assessors, and "explainers"
- Curate training data and follow accepted HSI design principles

Extend/develop T&E methods/tools to deal with stochastic, adaptive, emergent behaviors, and AS-specific vulnerabilities

- Methods/tools for complex, non-stationary, and non-deterministic systems
- Account for "emergent behavior" and defining the SUT
- New statistical engineering methods for T&E design and analysis
- Assessment/mitigation of subtle cyberattacks and adversarial attack vectors

Invest in infrastructure and process

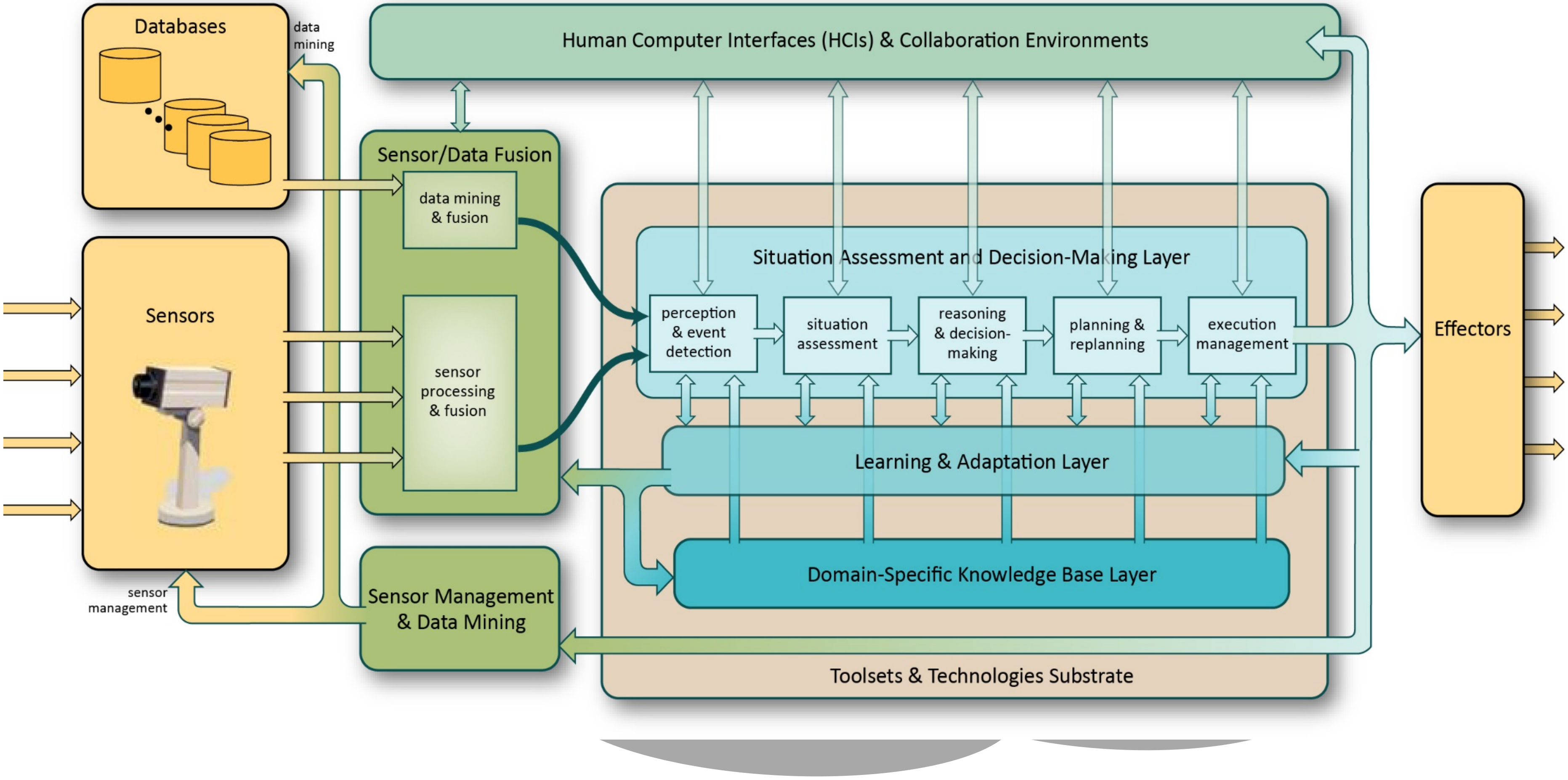
- Develop unifying infrastructure for requirements generation/traceability
- Move to "T&E Lifecycle" viewpoint and Invest in "digital modernization"
- Make massive use of M&S, test automation, & data analytics everywhere

Human-system teaming

- View the H-S Team as the SUT and embrace co-development of CONOPS with ASs

AS: Autonomous System

Common Framework for Autonomous Systems


Databases
data mining
Human Computer Interfaces (HCIs) & Collaboration Environments
Sensor/Data Fusion
data mining & fusion
Sensors
sensor processing & fusion
Situation Assessment and Decision-Making Layer
perception & event detection
situation assessment
reasoning & decision-making
planning & replanning
execution management
Effectors
Learning & Adaptation Layer
Domain-Specific Knowledge Base Layer
sensor management
Sensor Management & Data Mining
Toolsets & Technologies Substrate

What Would a Common Framework Buy Us?

- Provide common structures for many autonomous systems…
  - Internal component functions, their relationship to each other and the environment, and principles governing their design
- …to support parallel development efforts in different areas
  - Different groups can work complementary subsets of the problem, connecting with one another via the framework
- Develop unifying “science of autonomy” across 1000’s of “one-offs” now in the engineering community…
- …and point to where the S&T community needs to invest
  - Develop missing or inadequate functionalities
- Serve as foundation of an AS Open Systems Architecture (OSA)…
  - Encourage reuse of developed modules across applications
- …and support interoperability across DOD
  - eg, AF ISR UAVs cooperatively teaming with Navy attack UUVs