INSTITUTE FOR DEFENSE ANALYSES

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

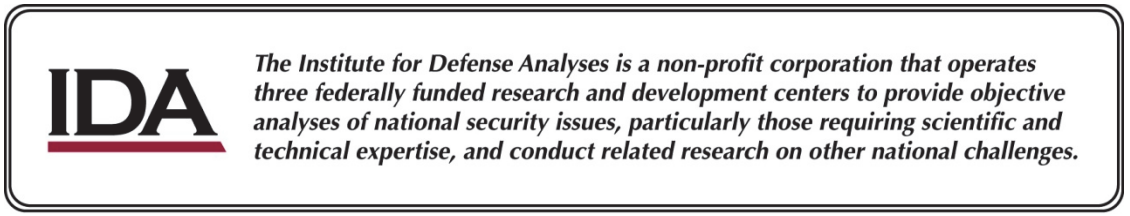
March 2019

Approved for public release.
Distribution is unlimited.

IDA Document NS D-9266

Log: H 2018-000389

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-229990, "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule, Director, and consisted of Stacey L. Allison and Stephanie T. Lane from the Operational Evaluation Division, and James M. Gilmore from the System Evaluation Division, David A. Sparrow and Poornima Madhavan from the Science and Technology Division, and Brian A. Haugh from the Information Technology and Systems Division.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

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

Executive Summary

The purpose of this briefing is to provide an executive-level overview of a more detailed, working-level framework for testing systems with autonomy (SWA). The brief outlines the challenges and broad-stroke reforms needed to prepare for next century's test challenges. The suggestions outlined here are not meant to be final.

A fundamental challenge in autonomy is developing trust in its decision-making capacity across all situations and environments it potentially will encounter. We can make many assumptions about human decision making that cannot be made about machine decision making. If we think of warfighters as a system being acquired, each individual unit has undergone decades of field testing regarding its quirks, and the entire manufacturing line has gone through tens of thousands of redesigns to increase efficiency. Developing human intelligence also takes a long time—decades of direct, time-intensive training from experts. Developing trust in that intelligence also takes time. If a person has survived to age 18, we can make inferences about their ability to navigate a plethora of complex, three-dimensional environments. For example, we're confident they can walk in sand, in snow, and on aspalt because we're confident they have a generalizable method of walking. We can't assume the same is true about autonomous systems—we need evidence.

To get this evidence, we would have to either use large tests or take a new approach to testing. Because larger tests are economically and politically infeasible, we have to reform the way we test. This is a solvable challenge, but it will require us to focus on building a "body of evidence" over time, with reforms touching every aspect of the testing lifecycle, including acquisition. We will have to adopt stricter criteria for acceptable testing. Finally, successful implementation of these new testing strategies will require a differently talented workforce.

A. Conceptual Framework for Autonomy

Autonomy comes in two flavors: procedural and executive. Procedural autonomy includes systems whose operationally relevant tasking ("should") decisions are made by humans, but there is some flexibility in how these systems pursue their given goal. They pick the "how" of a task. For example, a modern

missile may not have autonomy about what is being targeted but it may have autonomy in how it maneuvers to get to the target. Procedural autonomy contrasts with executive autonomy. A system with executive autonomy makes its own "should" decisions: Should I point my sensors at this location, should I shoot down this target, should I classify this person as a threat?

For a system with procedural autonomy, we don't really care why it makes the decisions it does. If it gets the job done across the operational space, that's really all that matters. But when you get into systems with executive autonomy, the operational space explodes. We do care why it makes the decisions it does. For example, did it classify that person as a threat just because they have a gun?

Some SWA already have executive autonomy. However, these systems all run off of symbolically coded logic. Aegis is one example. These systems might be complex, but their decision rules can still be interpreted by a human. When you get into systems whose rules were developed by a machine learning algorithm, suddenly we encounter situations where the "why" is not intuitive.

This broad framework of procedural versus executive autonomy is important, because even though the broad framework for designing a test is similar for both types of autonomy, there are critical distinctions pertaining to executive autonomy. We are still designing the test around a mission — that doesn't change. We still have to cover the operational space, and the things that could affect mission performance are still our test factors. Essentially, what changes is that mission performance becomes equivalent to whether the system made the right decision. For procedural autonomy, mission performance involves the same metrics we have traditionally measured, such as probability of kill. But for executive autonomy, we need some way of scoring whether the system made the correct decision, and we need to include as part of testing various factors that change what the correct decision is. Furthermore, we need to test what the mission would require, not what the system is limited to considering. For example, if an autonomous system cannot compute the likelihood of fratricide but the mission involves that risk, the testing scenarios should still let us uncover how likely that risk is.

Scoping tests will also look familiar. They should still be driven by the level of risk we are willing to accept for missing a problem or getting the wrong answer. They should still be driven by the severity of consequences in the mission and the size of the operational space. As the size of the operational space grows, we need to collect more evidence or we increase the risk of missing a problem. As the severity of consequences for the machine's decisions increases, we increase the significance of that problem.

B. The Basic Test Strategy

It is infeasible to create a brute force test that covers the entire operational space of autonomy. We must instead build a "body of evidence" over time, pulling from multiple data sources to answer our questions. We must use sequential testing, leverage modeling and simulation (while overcoming the unique challenges autonomy will pose), and eliminate the hard-line distinction between developmental testing and operational testing. Most importantly, systems must be designed to record data about themselves, by themselves. This cannot be an add-on for operational testing—it must be part of contractual requirements. Fortunately, this change will be an easier sell, as contractors will require this data infrastructure to successfully build and train the system anyway. If we are going to successfully test SWA in the future, the reforms must start now. Though implementation of these systems is relevatively far down the line, successful testing must modify the entire acquisition cycle, starting from conceptual development.

C. Reforming Test Standards for SWA

Autonomy will exacerbate current challenges in testing as well as introduce new problems. It will be important to raise the standards for measuring human-system interaction, determining what holes to accept in our coverage of the operational space, testing at the edges of the performance envelope, defining "operationally realistic," and handling cybersecurity testing.

Furthermore, new test methods will be needed for systems that make operationally relevant tasking decisions for themselves, systems that function as true teammates with a partner, systems that will continue to learn or adapt after fielding, and systems that have large operational spaces with catastrophic failure modes.

D. Recommendations

- Begin acquiring a workforce that has new skills in the domains of computer science, statistics and evaluation, and the human sciences.
- Advocate for Explainable Artifical Intelligence, testable requirements, and operationally driven requirements.
- Invest in range-realistic infrastructure for autonomy.
- View SWA testing as a hard problem, but one that can be solved if we start the reform effort ASAP.

Operational Testing of Systems with Autonomy

The absence of common sense prevents an intelligent system from understanding its world, communicating naturally with people, behaving reasonably in unforeseen situations, and learning from new experiences.
- David Gunning

Jane Pinelis, *Project Leader*

Chad Bieber, Heather Wojton, Daniel Porter,

Mike McAnally, Laura Freeman

10/18/2018

Setting Expectations

- This is an executive-level overview
  - Less deep coverage on many topics
  - A more detailed framework exists
- Focus is on operational test
  - But autonomy may need overall acquisition reform
- Brief is informational with some recommended COAs
- Our framework is not the final answer
  - Chief Scientist will iterate with us

An autonomous car shows up instead of a taxi

If all you knew was the AI had passed the same driving test a human currently takes, would you trust it?

That is the only information we have for human drivers.

Minimum requirements

Anybody can drive with Uber, although there are a few minimum requirements:

- Meet the minimum age to drive in your city
- Have at least one year of licensed driving experience in the US (3 years if you are under 23 years old)
- Have a valid US driver's license
- Use an eligible 4-door vehicle

We demand more evidence of machine capability

- Humans assumed to have other basic capabilities
  - Cannot assume machines have these
- Larger difference for autonomous weapons systems
  - DoD 3000.09
  - Warfighter perceptions
- Operational testing must provide assurance to warfighters that fielded autonomy is combat credible

Testing at the speed of relevance requires reform

Higher standards permit fewer holes in operational space

- These force us to have more evidence

Must move away from More Evidence = Bigger Tests

Change is needed to get evidence in new ways

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

A Solvable Problem

Autonomy: Making decisions based on environmental input

VS

Testing decision making is the key difference for autonomy

VS

We have tested some systems with autonomy

Our challenge will be systems with executive autonomy

- **Procedural:** How does the agent execute the task?
  - These are most of the systems we have already tested



- **Executive:** Should the task be executed?
  - Includes deciding what the task should be


With executive autonomy, we care about “why”

VS

Black Box

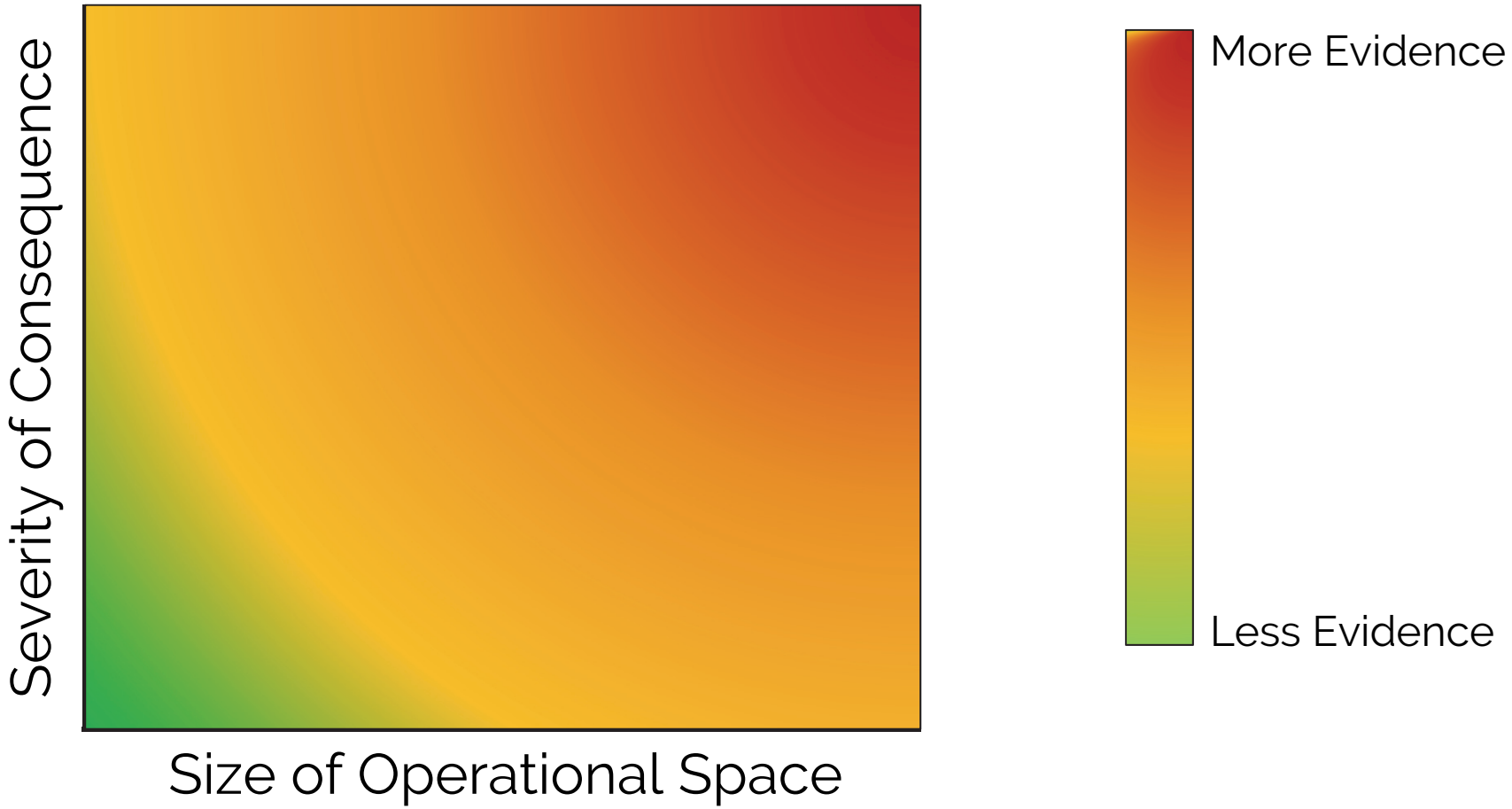
Especially for procedural autonomy, the test design process will remain broadly the same.

Testing should still be defined by the mission

vs

- Testers must cover the operational space
  - Mission requirements, not system limitations, should still drive the selection of OT&E factors
    - e.g., if AI cannot account for fratricide risk

Risk accepted should still drive test scope



Autonomy test methods already exist

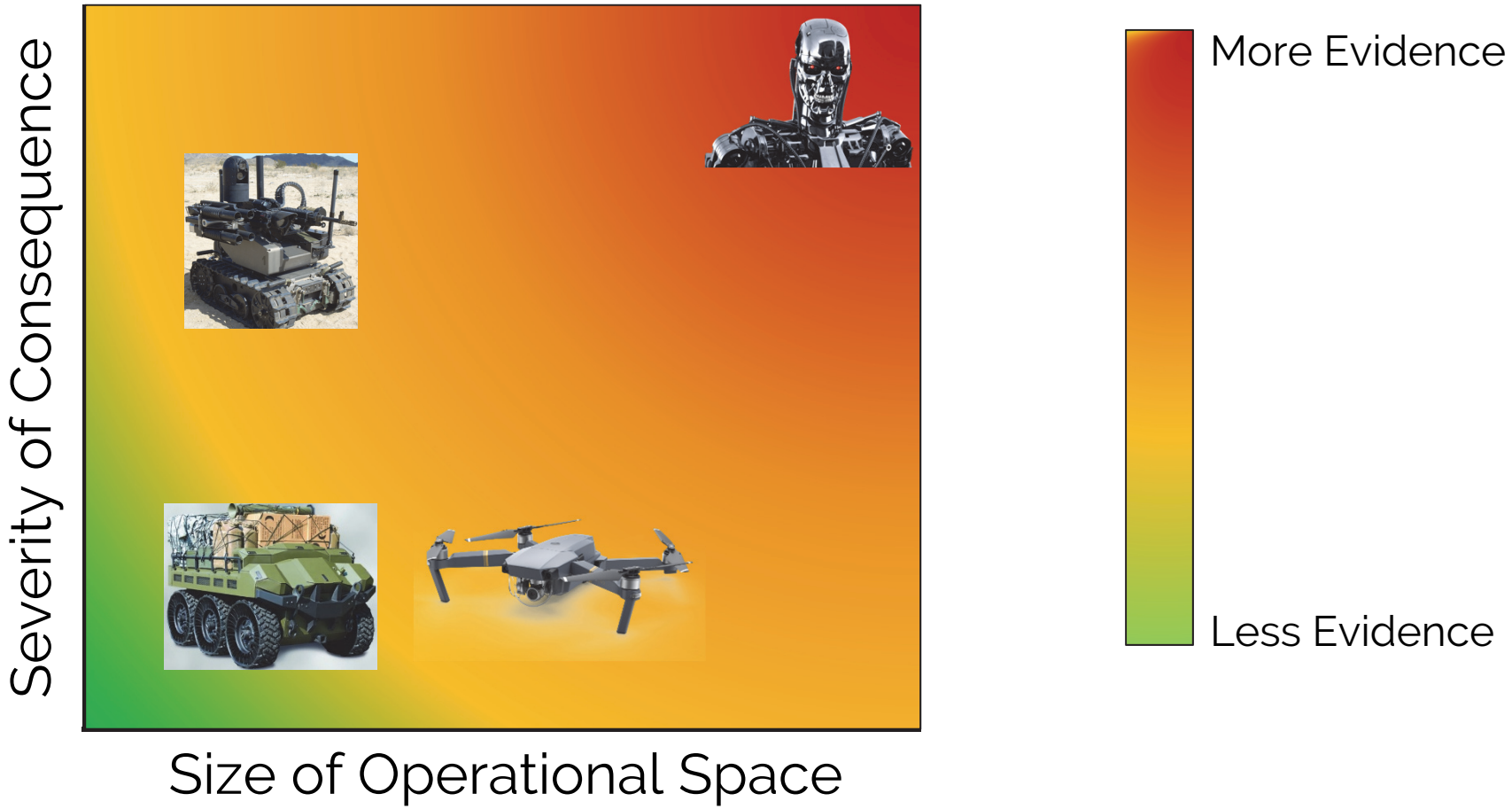
- **Challenge:** Build understanding of black-box decision-making engine

- **Solution:** Adopt methods from the human sciences
  - Humans really are black-box decisions makers
  - These fields built techniques for studying this challenge

Reforms needed for future systems must start now

- The systems we will test in 2-4 years may be okay
  - Methods not ideal, but problems not catastrophic
    - System decisions still mostly low risk

Reforms needed for future systems must start now



Reforms needed for future systems must start now

- The systems we will test in 2-4 years may be okay
  - Methods not ideal, but problems not catastrophic
    - System decisions still mostly low risk

- Changes needed must start at program inception
  - Systems where current method will fail are already being conceived
    - Won't be ready for test for years
  - Fixes to OT have to start before OT
    - Have to start now to be ready for systems in 10-20 years

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

Building a Body of Evidence While Preserving Timelines and Budgets

We must get more evidence without breaking budgets

- **Challenge:** Autonomy will require more evidence
  - More evidence doesn't have to mean more test points
    - Brute force testing would balloon budgets and timelines

- **Solution:** Build a "body of evidence" over time
  - **Targeted testing:** cover the space in intelligent ways
    - Each point must provide more evidential value
    - Focus on what test points allow us to learn about a system
  - Expand data sources that inform operational evaluation
    - More evidence without more test

Targeted testing must be informed by prior results

- Sequential testing guides targeted testing
  - Pick next test points based on what we learned in past
  - Test over time instead of one massive test
  - Helps maximize value of each point

- Modeling & Simulation can inform targeted testing

Autonomy will create new challenges for M&S

- Example Challenges:
  - Modeling target misidentification
    - Sufficiently model sensor inputs that lead to ID problems
  - Model resolution and fidelity
    - AI decisions will probably need better models than we currently use
    - Increases digital range demands
  - Resource investment
    - High up-front cost
    - Currently no universal architecture
  - Asynchronous processing
    - e.g., fusion of multiple sensors if one takes longer to process than the other

Targeted testing must not delay fielding

- **Challenge:** Sequential testing can expand timelines
  - ➢ Need to have previous test points to pick the next ones
    - Can't do this in a live test, so have to test over longer period

- **Solution:** Push the start of testing left
  - ➢ Begin collecting *operational-esque* data earlier

  - ➢ Earlier start means data must support both DT & OT
    - DT/OT needs to become a continuum
    - This is probably desirable for autonomy in any event

      AI needs realistic environment to see true behavior anyway

      OT needs to continue to enhance our understanding of system

Complex systems must expand sources of data

- **Challenge:** Some systems will just need more data
  - ➢ e.g., AI vs. Human Piloted Aircraft
    - Physical platform requires same amount of testing either way
    - The AI-pilot also has to be tested

      But compare to time spent on training/evaluating human pilots

- **Solution:** Leverage data from other sources
  - ➢ Data supporting FRP decision can come from beyond OT
    - Uncontrolled environments with operational flavor
    - e.g., training, late DT, early fielding, etc.

Data collection must be built in to the system

- To leverage other data sources, decisions and conditions must be recorded

- The system must record the data itself
  - Impossible to record data in many situations
  - Requires horde of observers when it is possible

- Data collection infrastructure must be a requirement

- This is not just for OT
  - Developers & DT will need the infrastructure too
    - Need to diagnose decisions to fix them

Reform must affect all acquisition

- Requirements
  - DOT&E should participate more actively
- Contracting
  - Access to algorithms and data
- System Design
  - Data recording infrastructure
  - Testable, traceable, and explainable
- Developmental Test
  - More operational realism
  - Integrated data needs with OT

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

OT Standards for AI & Autonomy

New standards for test adequacy must be developed

- Many systems do not need a unique framework
  - ➢ BUT Autonomy will exacerbate problems in testing
    - For autonomy, we must resolve currently accepted risks

- Some systems require methods we don't use yet
  1. Systems with executive autonomy
  2. Systems that are true teammates
  3. Systems that continue to learn after fielding
  4. Systems with large operational spaces with catastrophic failure modes

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

Current challenges exacerbated by autonomy

- Human-System Interaction (HSI) measurement
- Coverage of operational space
- Testing to system limits in operational environments
- Operational realism
- Adversarial vulnerability

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

Holes in coverage are riskier with autonomy

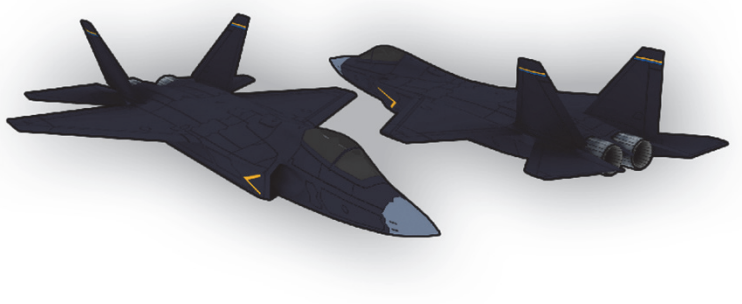
- We currently accept sparse coverage of space
  - Cost and we assume humans are flexible enough
- AI will lack this flexibility for the foreseeable future
  - Human supervision won't be sufficient
  - Unknown failure modes may have dire consequences
- **Challenge:** Adequately cover space in sequential test
- **Solution:** Develop test methods that combine space coverage with testing over time

We recommend you be the demand signal for this development

Test points should move toward performance limits

- Most current testing examines centroid of performance
  - Sometimes this is adequate
- Autonomous systems will have to be pushed
  - AI info processing may hit real-time limits more easily
- Problems disguised in the center of the envelope will emerge in full when AI is pushed to processing limits
  - We should place more focus on these edges

Operationally representative assets are a challenge



The OpRep challenge can be overcome

- It is possible under some designs to teach AI to pretend
  - Long, academic explanation for how
- Pretending is just forwarding a different tag
  - If Perceive_ID == “CONEX” then Decision_ID = “T-72”
- Allows us to test TTPs without real assets

Autonomy vulnerabilities go beyond cyber

- Training data manipulation
- Adversarial environment manipulation
- Direct cyber threats

Many autonomous systems would have adequate tests if these challenges were addressed.

Some autonomous systems will need new test approaches due to the type of autonomy they have.

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

New methods will be needed for:

1. Systems with executive autonomy
2. Systems that coordinate decisions with other agents
3. Systems that continue to learn after fielding
4. Systems assigned risky, complex/multipart tasks
   - Large operational space
   - Failure modes are catastrophic

Executive autonomy requires testing decision making

- We must test the quality of decisions made
  - ➢ Explicitly test situations that change the correct decision
  - ➢ Explicitly score quality of decision
  - ➢ Requires having both “should” and “should not” scenarios

Case Study

Teammate tests must disentangle contributions

- Many systems aren't teammates, just fancy tools
  - ➢ Coordinated decisions are hallmark of teammates
- **Challenge:** Who is responsible for success or failure?
- **Challenge:** Does it team well with different people?
- **Solution:** Measure the performance of multiple partner-pairs on the same mission

Case Study

Systems that continuously learn must be recertified

- **Challenge:** A system that adapts after fielding may learn bad habits or may have degraded performance
- **Challenge:** System not "production representative"
- **Solution:** Regularly recertify AI skills

Case Study

Risky, complex systems should not be fielded all at once

- **Challenge:** Testing systems with large operational spaces where failure risks human life
  - ➢ Too many opportunities for holes in coverage to lead to catastrophic consequences

- **Solution:** Limited or Incremental Capability Fielding
  - ➢ Complex tasks can be broken down into smaller ones
  - ➢ Choose a subtask with acceptable risk and test that
    - ▪ If it passes this test, approve it for fielding on that task
      - ❑ Potentially limit to human-supervision
  - ➢ Collect field data through built-in infrastructure
    - ▪ Over time adjust risk of approved tasks

Case Study

Key Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
  - Data recording infrastructure must be part of system
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

The methods exist, but are we ready?

Our workforce must adapt to be ready for autonomy

- Autonomy T&E will require expertise in:
  - ❑ Computer Sciences
    - Difficult to recruit and retain for DoD
  - ❑ Statistics & Evaluation
    - Sophisticated analyses needed exceed current skill level
  - ❑ Human Sciences
    - Currently <2% of OTA workforce has any background
- Workforce of tomorrow must expand these capabilities to have credible autonomy evaluation

You can advocate for COAs related to autonomy

- Design systems and write requirements as testable
  - Ensure data collection infrastructure is part of system
  - Advocate for Explainable AI
    - Testable, traceable, and explainable decisions
- Invest in range-realistic infrastructure for autonomy
  - Computational demands will outstrip our capacity for DT
  - Realistic, current threat environments
- Start building DoD's talent pool

Takeaways

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require change
  - Focus on building "body of evidence" over time
- New test standards must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

Thank you

Backup Slides

Cybersecurity Testing Placeholder

AWS will be held to highest standard

BLUF

- This is a hard problem but it is not impossible
  - We already test systems with autonomy
- Credible testing without larger tests will require reform
  - Focus on building "body of evidence" over time
- New standards for test adequacy must be developed
  - "Adequate test" now won't be for advanced autonomy
  - Have to adopt new methods for some system types
- Challenges of autonomy will require workforce change
  - Computer, statistical, and human sciences needed

This higher standard is codified in DoD policy

- DoD Directive 3000.09
  - 4.c.(3): "Autonomous weapon systems may be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against materiel targets…"
  - 4.d: "Autonomous or semi-autonomous weapon systems intended to be used in a manner that falls outside the policies … must be approved by the Under Secretary of Defense for Policy (USD(P)); the Under Secretary of Defense for Acquisition, Technology, and Logistics (USD(AT&L)); and the CJCS before formal development and again before fielding…"

For some systems, current methods won't give us the type of data we need to judge effective, suitable, or survivable.

DOT&E is responsible for T&E standards for autonomy

- 3000.09 Enclosure 4
- 4. DOT&E. The DOT&E shall:
  - a. Provide **principal oversight responsibility for the development of realistic operational T&E standards for semi-autonomous and autonomous weapon systems**, including standards for T&E of any changes to the system following IOT&E, in accordance with subparagraph 4.a.(1) above the signature of this Directive and Enclosure 2.

  - b. Evaluate whether semi-autonomous and autonomous weapon systems under DOT&E oversight have met sufficient V&V and T&E in realistic operational conditions, including potential adversary action, in order to minimize the probability and consequences of failures that could lead to unintended engagements or to loss of control of the system to unauthorized parties.

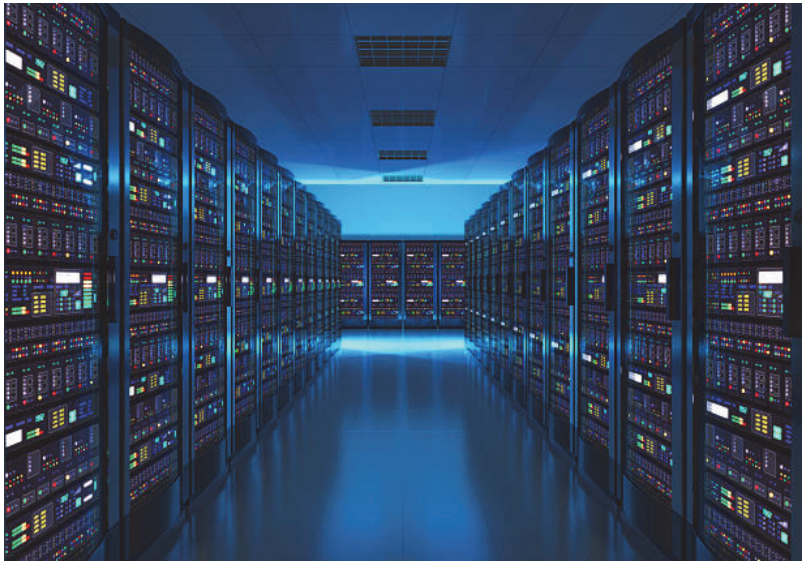
OT solutions require changes to design and acquisition

- For these solutions to work, we also have to:
  - **Have test infrastructure built in to the software**
    - Not an add-on for OT. Developers and DT will need it anyway
    - This requirement is critical
  - Testers have to "see under the hood" of AI
    - System architecture may change how it should be tested
  - Systems must be understandable by humans
    - DARPA Explainable AI
    - This is a design goal that should be tested
    - Explainable AI is better for use but also easier to test

DoD Data Infrastructure Gap

Need data information systems that:

- Hold **large** amounts of data
  - 24-hour High Definition video from 1 camera is ~ 1 TB of data
- Are accessible for high-volume upload and download
  - Netflix communicates 3 GB data in a 1-hour High Definition video stream
  - Streaming data to pull features will require high bandwidth
- Follow best practices on Prevent-Mitigate-Recover (PMR)



*A current focus of the Joint Enterprise Defense Infrastructure (JEDI) Program

OT&E of autonomy is a solvable problem

- Designing AI will be hardest
- Developmental testing for AI demands innovation
- OT requires change but not revolutionary methods

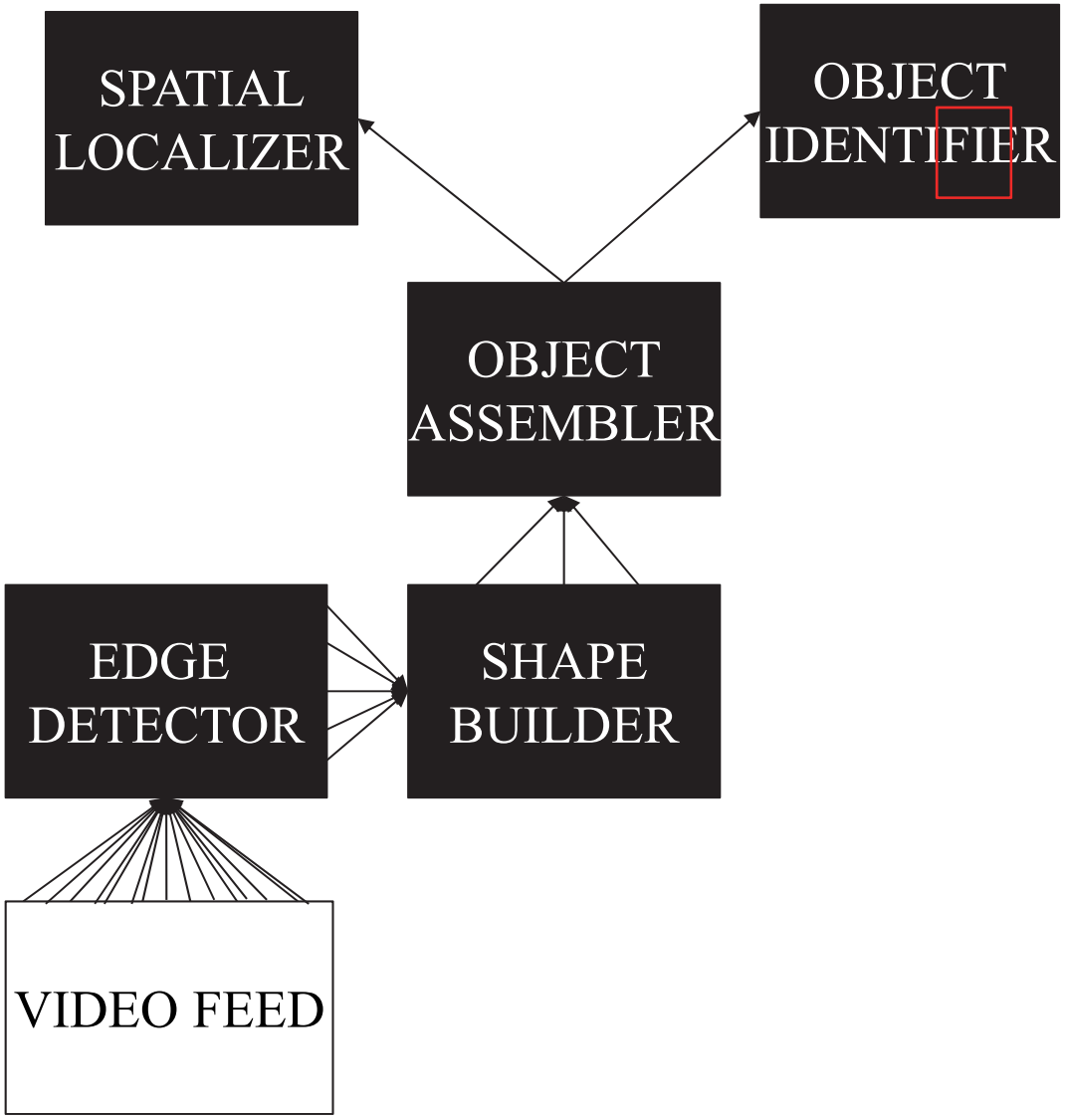
We need to reform contracting

- To test the system correctly, we must understand it
  - Already we have tested systems with autonomy where companies refused to share details of their algorithms

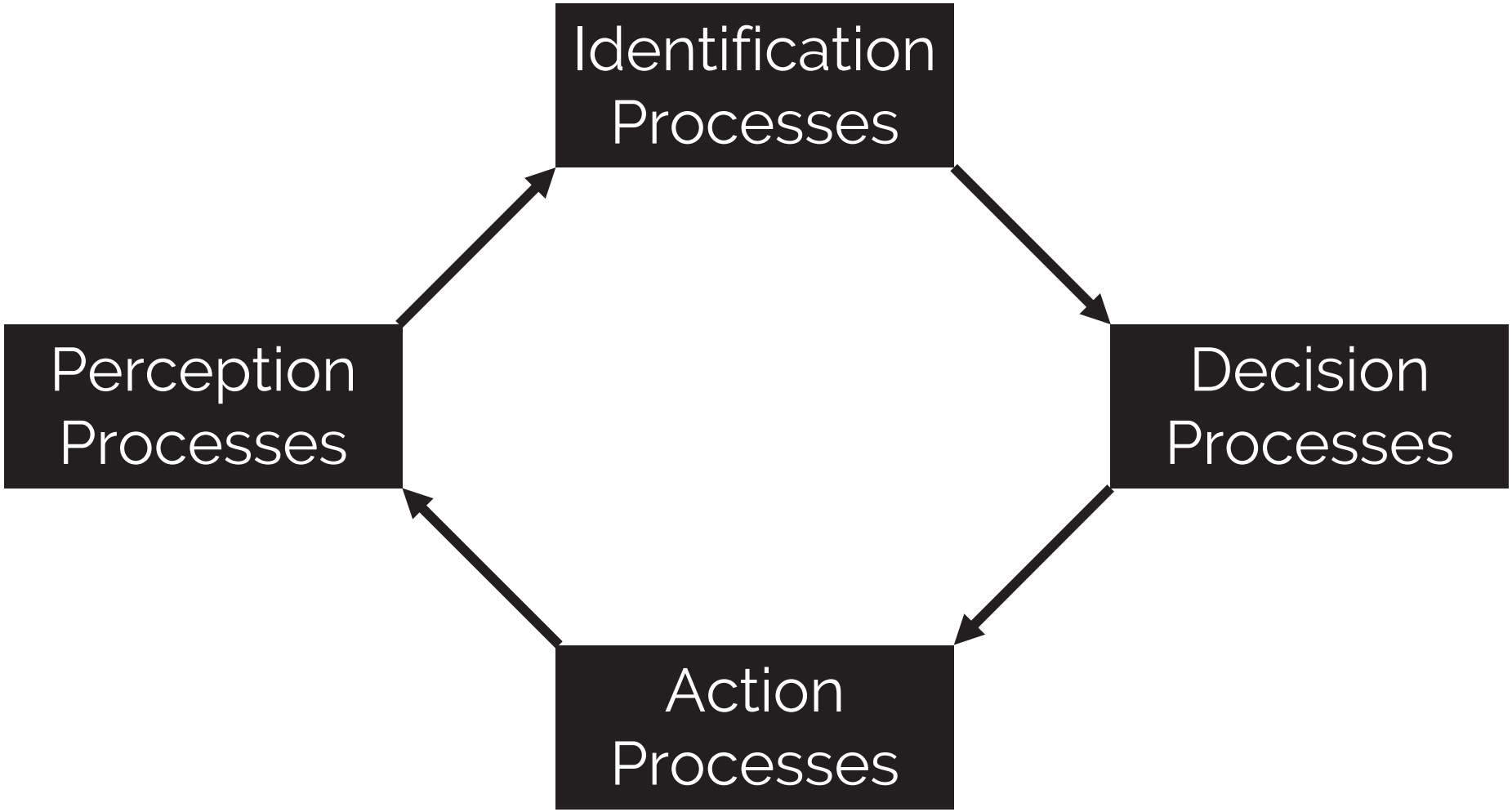
Autonomy is less “black-box” than many claim

Sophisticated autonomy will be networked black boxes

Task: Identify what things are where



Different processes support OODA stages



Solution: Test stages separately

- Perceive → Identify
  - Does Detect, Recognize, & Identify happen correctly?
  - If we are avoiding real assets for safety or expense
    - Test only DIR processes under realistic conditions
    - No actions taken, maximize asset time used for DIR
  - If we have no assets, do our best to get real sensor recordings
    - Relies on intel community
    - If we can't get this, system needs to rely on logic

      This can still be tested

Different processes support OODA stages

Identification
J-31

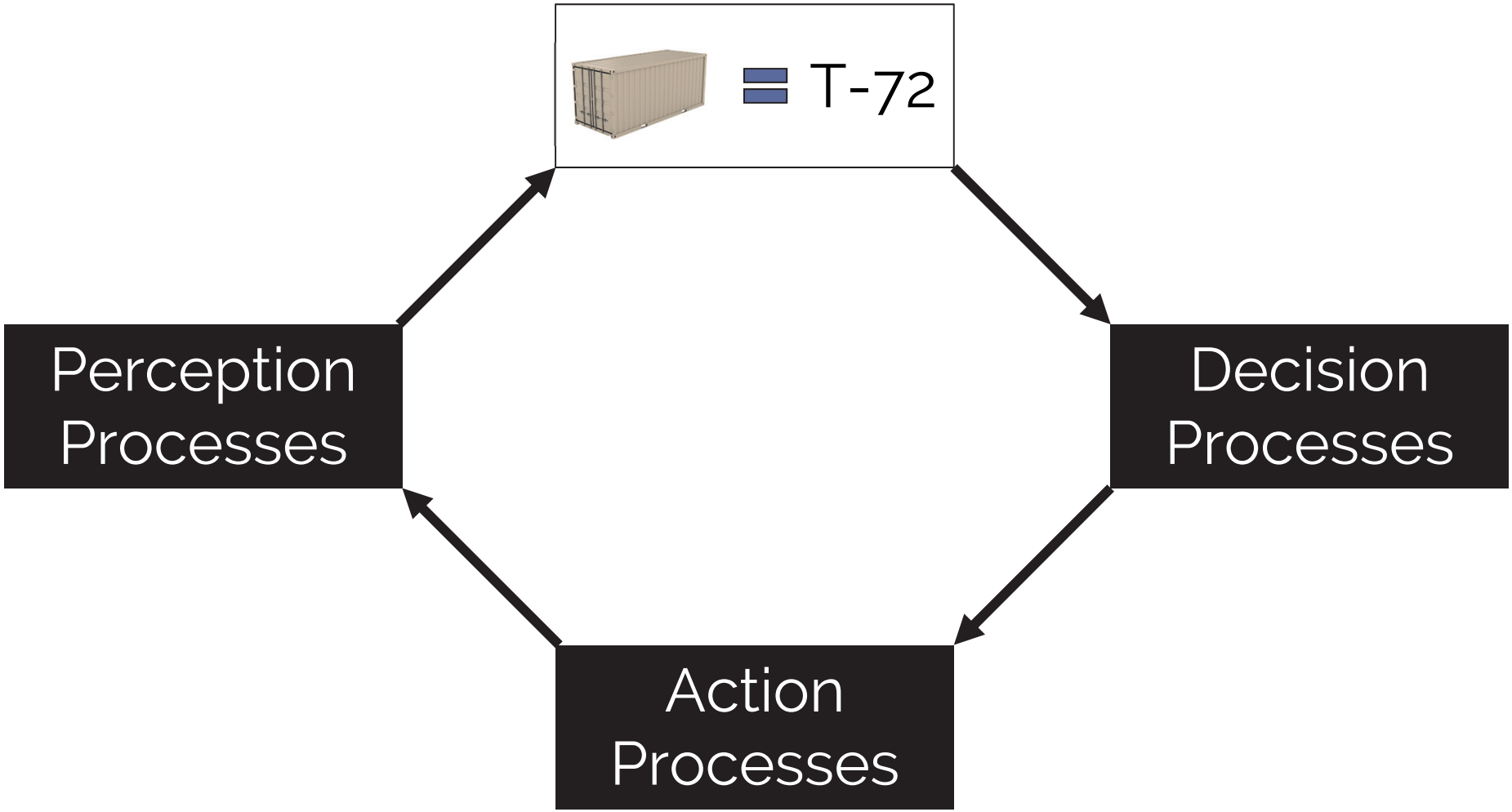
Decision
Processes

Action
Processes

Solution: Test stages separately

- Decision → Act
  - Does it choose the correct TTPs for a given situation?
  - We do not have to use real assets for all decision tests
  - Teach the AI to pretend like a human

Solution: Teach robots to imagine



323 TES Commander
@IanKnight35

Follow

What do you get when you combine the @LockheedMartin #F35, the @BoeingDefense #GBU39 and 1 awesome OT squadron? Unstoppable all weather accuracy with low collateral damage! First ever #F35 SDB employment: executed by the @Kon_Luchtmacht. #323TES #flexibilityiskey #candomentality

1:45 PM - 4 Oct 2018

Test Concept: Autonomous CAS

- Choose factors that should influence decisions, e.g.,
  - Potential for collateral damage/fratricide
  - Difficulty of target discrimination
  - Time pressure

- Explicitly score quality of decision, e.g.,
  - Objectively measure collateral damage
  - Subjectively rate if Ground Commander's Intent met

- Include "should" and "should not" scenarios
  - Test Type 3 CAS and see what choices it makes
  - Provide it 9-lines that are OBE to see if it still prosecutes

Test Concept: Autonomous CAS

- Over time, system gets better at CAS and molds to specific partners

- Every X months, unit performs recertification test
  - Analogous to training requirements for human pilots
  - Establish thresholds for cleared flight

- Built-in test infrastructure records data and sends back
  - Allows program and oversight to continue to monitor

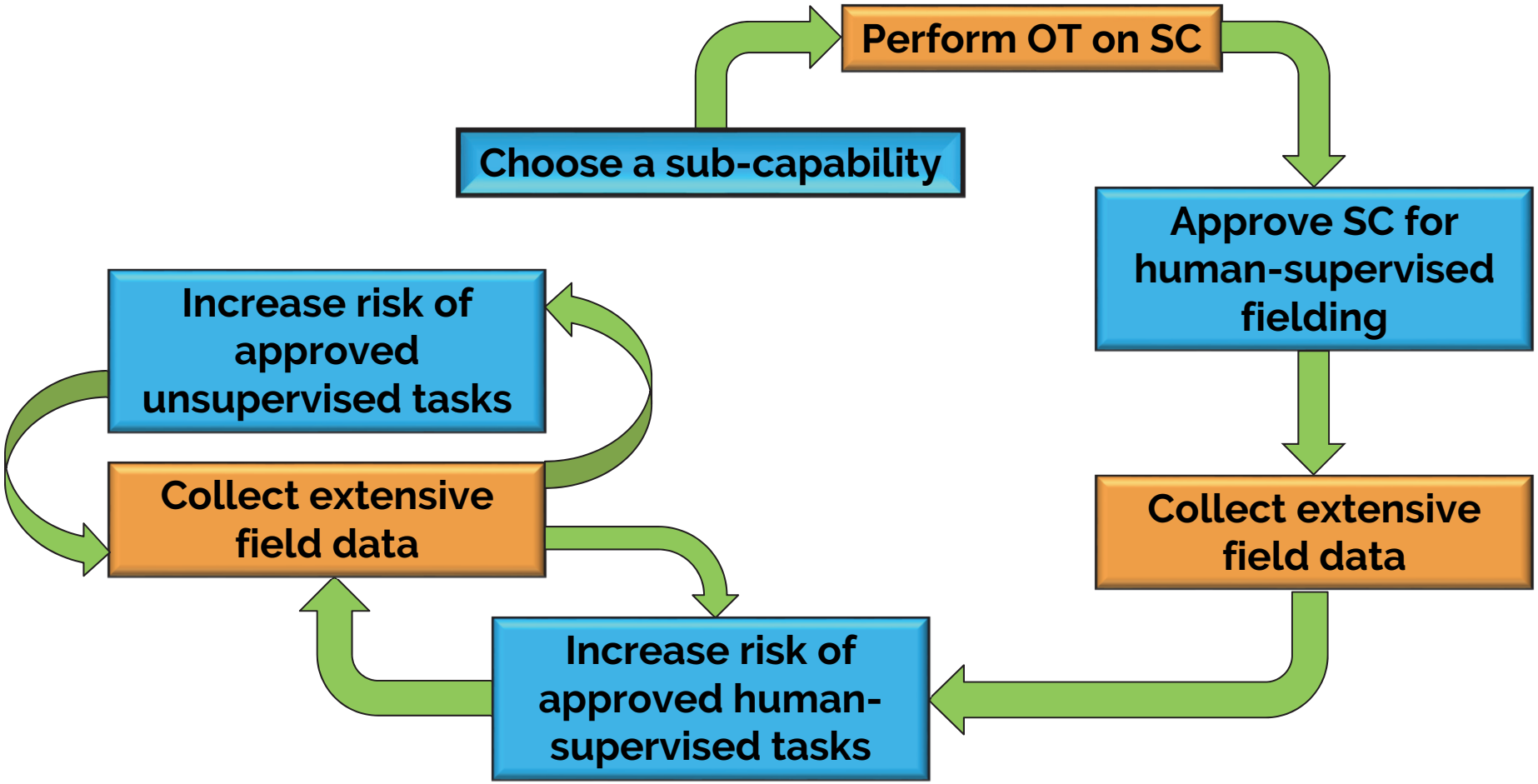
Test Concept: Autonomous Multi-role Fighter

- AI-piloted F-35 must perform many different missions
  - SEAD, CAS, DCA, Strike, etc., etc., etc.

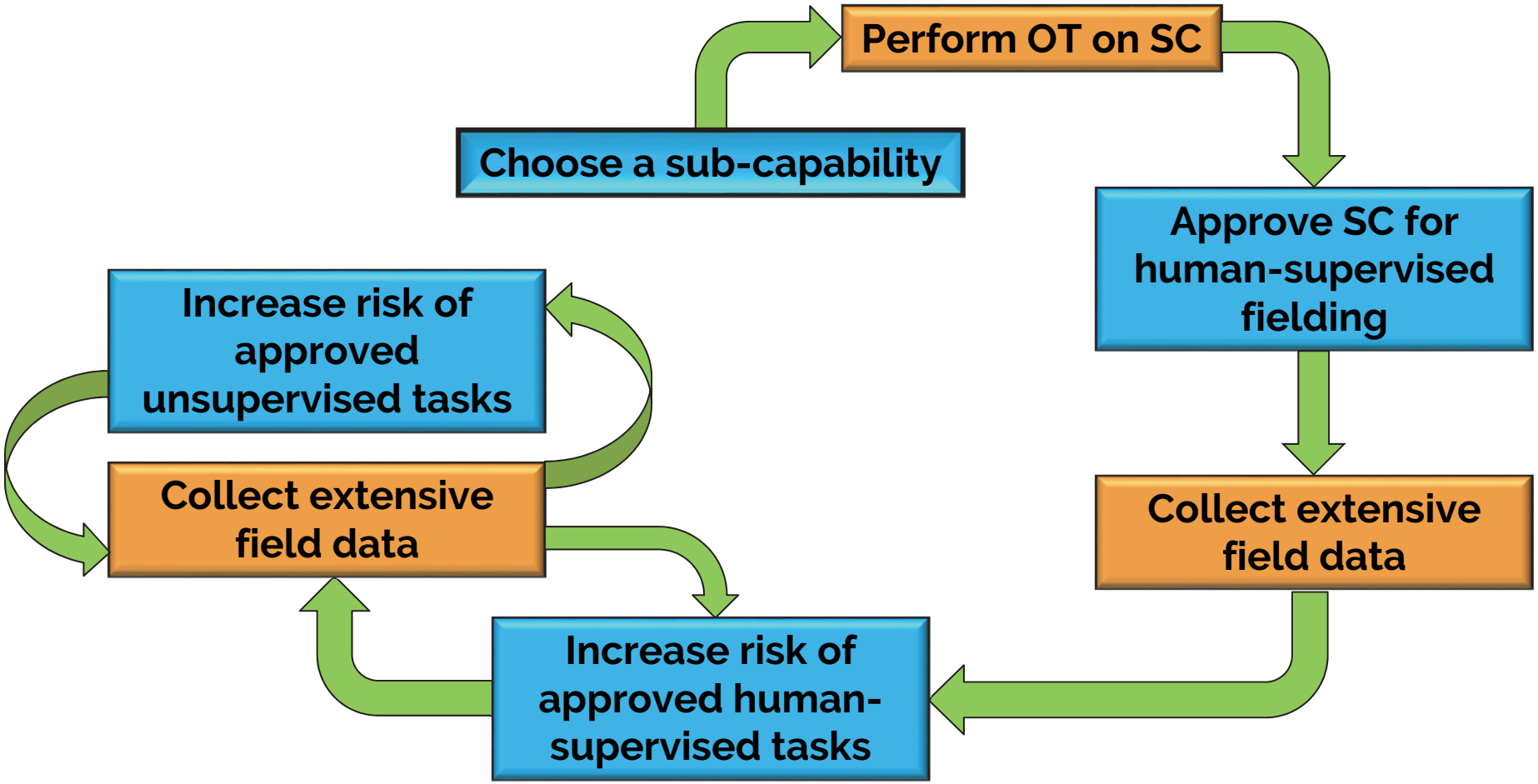
- Performing EA during a SEAD mission least risky
  - During training or in real ops, it is permitted to do that
  - Allow it to request to drop ordinance on materiel target
    - These data are automatically recorded by test infrastructure
  - Not permitted to act on, but have it record decisions it *would* make about attacking human targets

- Build data set of observed conditions and decisions
  - Evaluate confidence in its decisions based on these data
    - Approve it for unsupervised materiel destruction
    - Evaluate supervised human targeting

Risky, complex systems should not be fielded all at once



Risky, complex systems should not be fielded all at once



Test Concept: Autonomous CAS

- Over time, system gets better at CAS and molds to specific partners

- Every X months, unit performs recertification test
  - Analogous to training requirements for human pilots
  - Establish thresholds for cleared flight

- Built-in test infrastructure records data and sends back
  - Allows program and oversight to continue to monitor

A Framework for Designing an Autonomy Test

1. Define the mission
   - What tasks are required to complete the mission?
   - What inputs affect decision making for each task?
     - These become your factors
   - What are the consequences of the tasks?
     - This should drive test size
2. Define the autonomy
   - Does it have executive or procedural autonomy?
     - If executive, include “should not” conditions
   - Is it a special case? (Teaming, Learning, CCV)
     - Teaming: Matched Pairs
     - Learning: Recertification Testing
     - CCV: Limited Capability Fielding

Test points should move toward edges of the envelope

- Most current testing examines centroid of performance
  - Not always bad idea not to push to limits
- Autonomous systems will have to be pushed
  - Brain better suited to chaotic environment than computer when it comes to complex tasks
    - Parallel continuous coding vs. Serial binary processing
  - AI info processing will hit real-time limits more easily
    - Easier for adversaries or environment to break AI OODA loop once we move beyond narrowly scoped autonomy
- Problems disguised in the center of the envelope will emerge in full when AI is pushed to processing limits
  - We should place more focus on these edges

Risky, complex systems should not be fielded all at once

- **Challenge:** Testing systems with large operational spaces where failure risks human life
  - Too many opportunities for holes in coverage to lead to catastrophic consequences

- **Solution:** Limited or Incremental Capability Fielding
  - Complex tasks can be broken down into smaller ones
  - Choose a subtask with acceptable risk and test that
    - If it passes this test, approve it for fielding on that task
      - Potentially limit to human-supervision
  - Collect field data through built-in infrastructure
    - Over time adjust risk of approved tasks

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
March 2019	OED Draft	

4. TITLE AND SUBTITLE

Operational Testing of Systems with Autonomy

5a. CONTRACT NUMBER

HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Daniel Porter (OED); Chad Bieber (OED); Heather Wojton (OED); Laura Freeman (OED); Michael McAnally (OED); Yevgeniya "Jane" Pinelis (OED)

5d. PROJECT NUMBER

BD-9-2299

5e. TASK NUMBER

229990

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

D-9266-NS

H 2018-000389

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Heather M. Wojton (OED)

14. ABSTRACT

Systems with autonomy will pose unique challenges for operational test. This document provides an executive level overview of these issues and the proposed solutions and/or reforms. In order to be ready for the testing challenges of the next century, we will need to change the entire acquisition life cycle, starting even from initial system conceptualization. This briefing was presented to the Director, Operational Test & Evaluation along with his deputies and Chief Scientist.

15. SUBJECT TERMS

acquisition reform; AI; artificial intelligence; autonomy; test concept; TEV&V

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Heather M. Wojton
Unclassified	Unclassified	Unclassified	Unlimited	93	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6811