INSTITUTE FOR DEFENSE ANALYSES

Demystifying the Black Box: A Test Strategy for Autonomy

Heather M. Wojton, Project Leader

Daniel J. Porter

April 2019

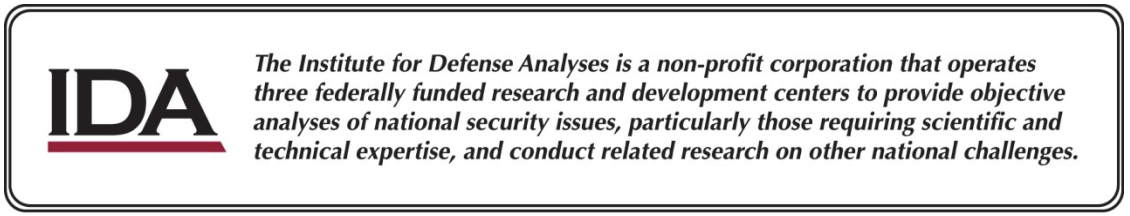
Approved for public release.
Distribution is unlimited.

IDA Document NS D-10465-NS

Log: H 2019-000126

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

Approved for public release; distribution is unlimited.



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
Review was conducted by Robert R. Soule, Director, and Heather M. Wojton and Rebecca M. Medlin from the Operational Evaluation 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-10465

Demystifying the Black Box: A Test Strategy for Autonomy

Heather M. Wojton, Project Leader

Daniel J. Porter

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

1. Executive Summary

The purpose of this briefing is to provide a high-level overview of how to frame the question of testing autonomous systems in a way that will enable development of successful test strategies. The brief outlines the challenges and broad-stroke reforms needed to get ready for the test challenges of the next century.

The presentation attempts to convince the audience of three things. In order of importance, they are: (1) the goal of testing should be developing and confirming a generalizable model of system decision-making, (2) the test lifecycle must be a continuum, not discrete categories operating as independent fiefdoms, and (3) the fundamental challenge of testing autonomy and AI is generalizing to unobserved situations. As the other points flow from the third, we will discuss it first.

A. Generalizing to Unobserved Situations

A fundamental challenge in autonomy is developing trust in its decision-making capacity across all situations and environments it will encounter. We are able to make a lot of assumptions about human decision-making that we cannot make for machines. We are able to do this because we have a mental model of how humans make decisions—a model that includes global Goals that we trust, and a moment-to-moment decision Process that we view as reasonable. Furthermore, any adult human has spent decades doing operational testing of the Capabilties on which this decision model depends.

We have no similar understanding of the Goals, Process, or Capabilities of a "black box" autonomous system. We fear discontinuities in its decision-making; that is, we worry that it is not actually considering the information we would deem most relevant (e.g., the location of the edge of the road) but instead is processing correlated but not globally helpful information (e.g., the presence of a sidewalk).

When we test physical systems, we can test relatively few points because we have an underlying model that allows us to interpolate performance between observations. The more of a black box a system is, the less well it is modeled, less interpolation we can do, and the more data points we need. Conversely, the better the model, the fewer data points we need.

Approved for public release; distribution is unlimited.

B. Testing to Find a Model

Unless we want to use brute force testing and cover large operational spaces, one of the goals of testing must be to uncover the underlying decision model the system uses. This will enable interpolation. Discovering the model is easier for some types of autonomy than others.

Autonomy comes in two flavors: procedural and executive. Procedural autonomy includes systems whose operationally relevant tasking ("should") decisions are made by humans, but which have some flexibility in how they pursue their given goal. They pick the "how" of a task. For example, a modern missile may have autonomy in how it maneuvers to get to a target, but not what is being targeted. Procedural autonomy stands in contrast to "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?

It does not take much testing to discover the model for a system with procedural autonomy. It will either be an explicit physics model, or performance testing alone will make the model clear. However, systems with executive autonomy will challenge our current test paradigm, and we need to understand their decision models.

C. The Lifecycle of Test

To evaluate the capabilities of autonomous systems to an acceptable level of risk, it will require more data. Therefore, we must find more efficient ways of testing or be willing to increase test budgets and delay programs. Part of seeking improved efficiency is attempting to understand the decision model. This is much easier if the decision process is designed from the beginning as a cognitive architecture. This is also made easier through the use of "targeted testing," which is a collection of methods such as design of experiments and sequential testing that aim to improve the evidential value of each formal test point. If targeted testing is going to be successful, however, then it must be approached in a unified way throughout the entire testing lifecycle. Otherwise, timelines and test inefficiencies will defeat the effort. To this end, we recommend that contractor, developmental, and operational testing be treated less like independent fiefdoms, and more like a continuum of tests answering different aspects of the same question.

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Demystifying the Black Box: A Test Strategy for Autonomy

Dr. Daniel Porter

Institute for Defense Analyses

11 April 2019

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Talk Takeaways – Order of Importance

1. The goal of testing should be developing and confirming a generalizable model of system decision-making.

2. The lifecycle of test must be a continuum, not discrete categories operating as independent fiefdoms.

3. The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations.

Approved for public release; distribution is unlimited.

Talk Takeaways

The goal of testing should be developing and confirming a generalizable model of system decision making.

The lifecycle of test must be a continuum, not discrete categories operating as independent fiefdoms.

The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations.

Approved for public release; distribution is unlimited.

Talk Takeaways – Order of Discussion

A. The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations.

B. The goal of testing should be developing and confirming a generalizable model of system decision-making.

C. The lifecycle of test must be a continuum, not discrete categories operating as independent fiefdoms.

Approved for public release; distribution is unlimited.

The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations.

An autonomous car shows up instead of a taxi

Approved for public release; distribution is unlimited.

If the AI certification process were just the same road test that humans take, would you trust it?

Approved for public release; distribution is unlimited.

Neither human nor machine has encountered all of the driving situations that it likely will, but we trust the human because we have a model of his or her decision process, which itself is operationally tested by living.

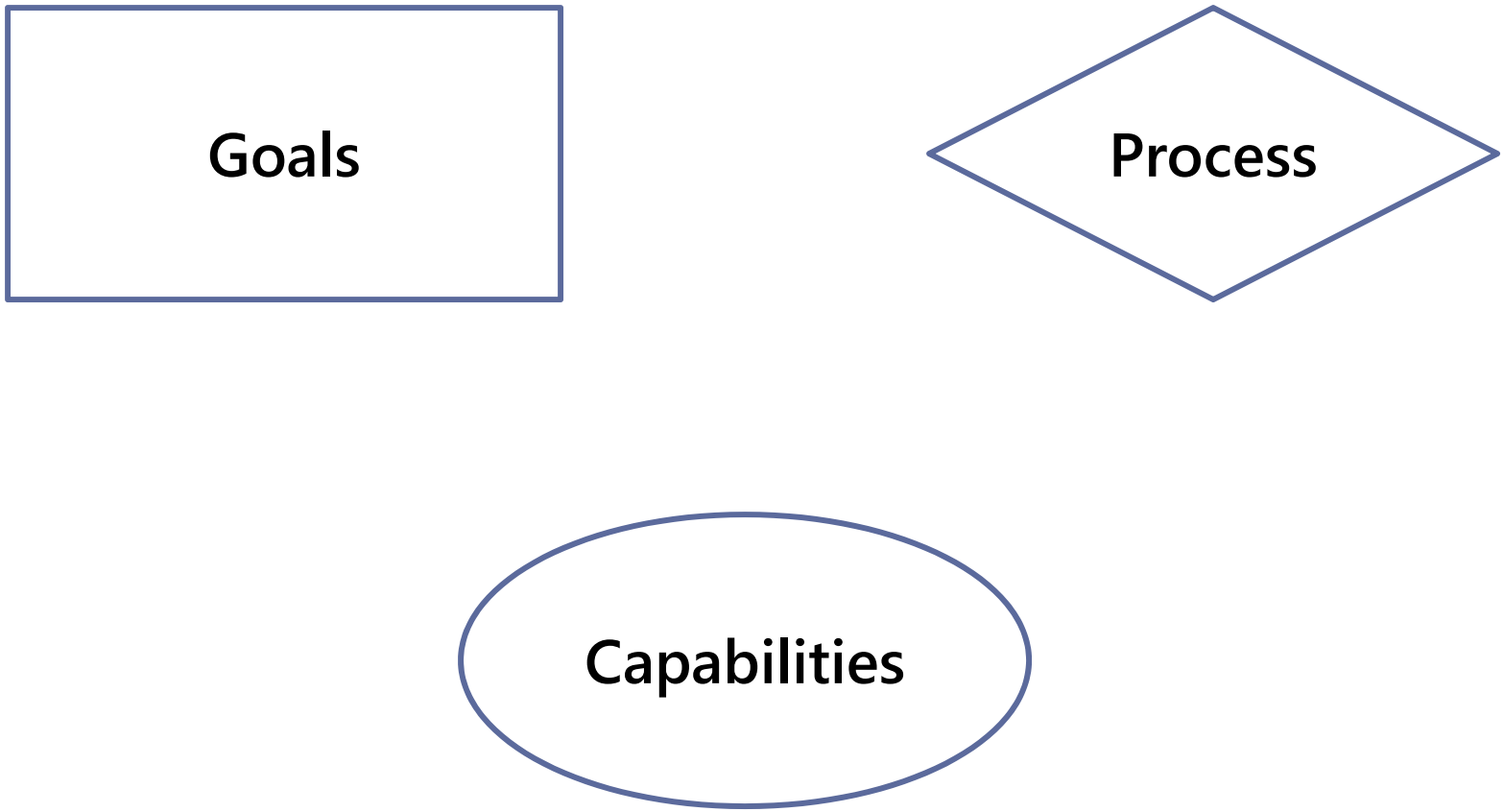
Approved for public release; distribution is unlimited.

What do I mean by model?

Are you the sort of man who would put the poison into his own goblet or his enemies? Now, a clever man would put the poison into his own goblet because he would know that only a great fool would reach for what he was given. I am not a great fool so I can clearly not choose the wine in front of you...But you must have known I was not a great fool; you would have counted on it, so I can clearly not choose the wine in front of me.

Approved for public release; distribution is unlimited.

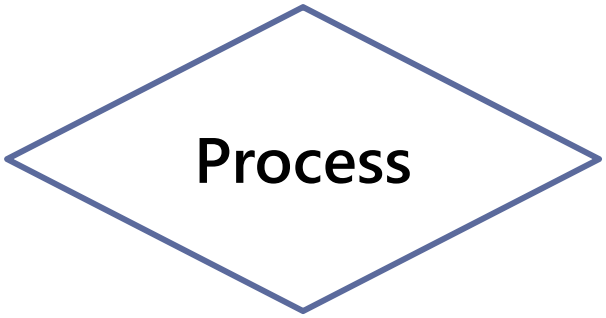
Trust of decision making has three basic inputs



Approved for public release; distribution is unlimited.

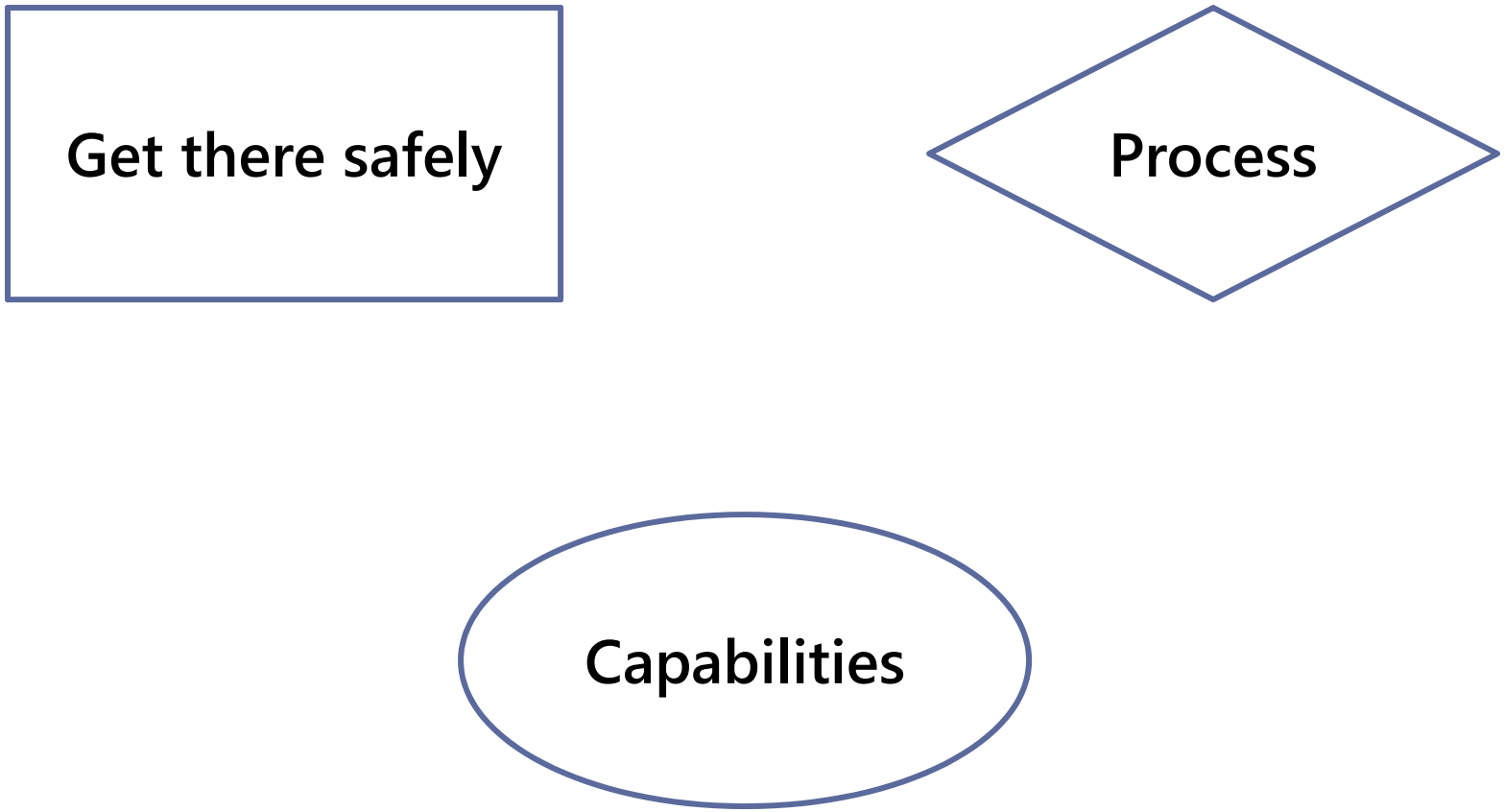
Trust of decision making has three basic inputs

- Don't die
- Don't get arrested
- Get paid



Approved for public release; distribution is unlimited.

Trust of decision making has three basic inputs



Approved for public release; distribution is unlimited.

Trust of decision making has three basic inputs

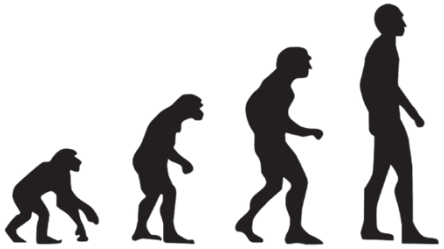
Get there safely

Capabilities

Approved for public release; distribution is unlimited.

Human capability is well-tested

Manufacturing Line



Quality Assurance

Approved for public release; distribution is unlimited.

Why don't we trust the machine?

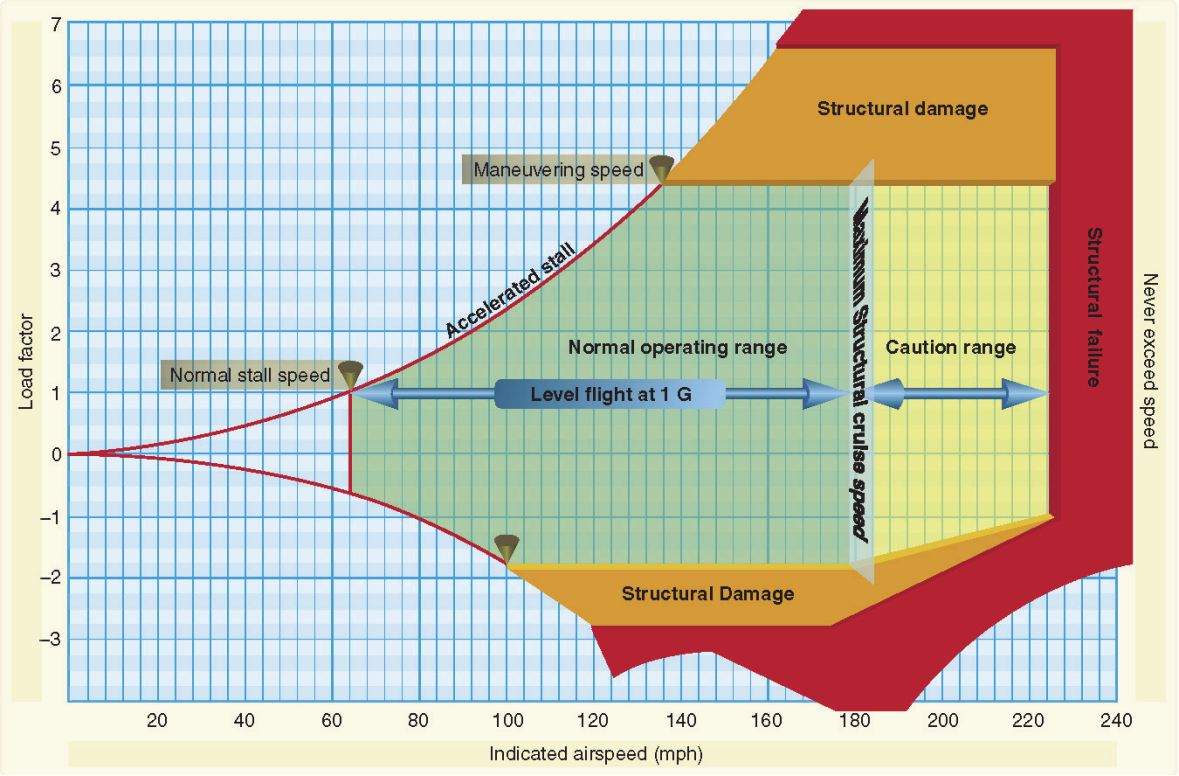
- If decision engine is a black box, we don't understand:
  - Global **Goals**
  - Moment-to-moment decision **Process**
  - Underlying **Capabilities** those processes depend on

- People fear discontinuities in decision-making
  - I think it uses X to make decisions, but it really uses Y
    - Y is highly correlated with X, and tested points confound them
      - E.g., Doesn't track edge of road, just tracks sidewalks

Approved for public release; distribution is unlimited.

The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations

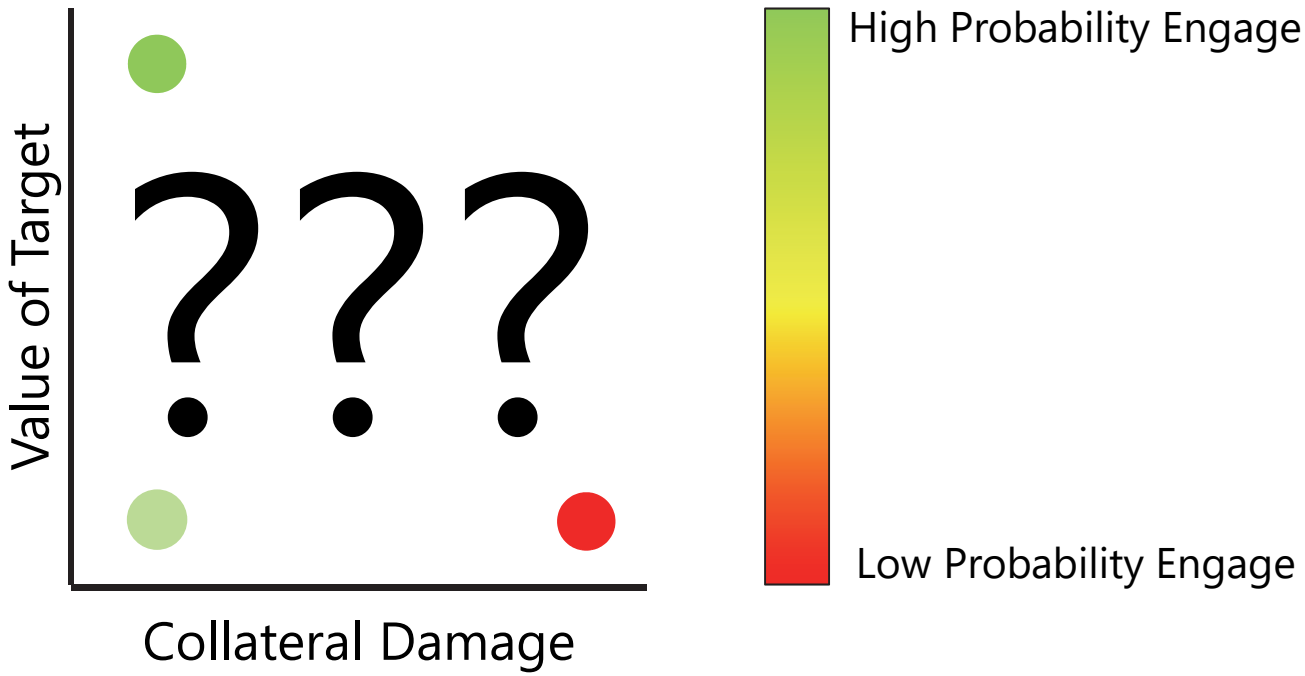
- Aerodynamics model allows inference



Approved for public release; distribution is unlimited.

The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations

- We don't have models of AI decision-making



Approved for public release; distribution is unlimited.

The goal of testing should be to develop a generalizable model of system decision-making.

Testing is about assurance

- Does the system meet its requirements?
  - Contractual
  - Operational

  } Hopefully aligned

- To what extent do different factors affect performance?
  - Identify areas for improvement

  - Inform development of TTPs

Approved for public release; distribution is unlimited.

Testing is about assurance

How do field commanders have assurance that the system they have is appropriate for the situation they face?

Approved for public release; distribution is unlimited.

Testing is about assurance

1. The systems Goals & Processes are reasonable.

2. The Capabilities on which those depend function.

Approved for public release; distribution is unlimited.

There are two broad test approaches

- Brute Force

  - ❑ Cover operational space sufficiently for acceptable level of risk
  - ❑ Black box forces this approach

- Interpolation

  - ❑ Observe limited points and predict between observations
  - ❑ Having underlying model enables this approach

Approved for public release; distribution is unlimited.

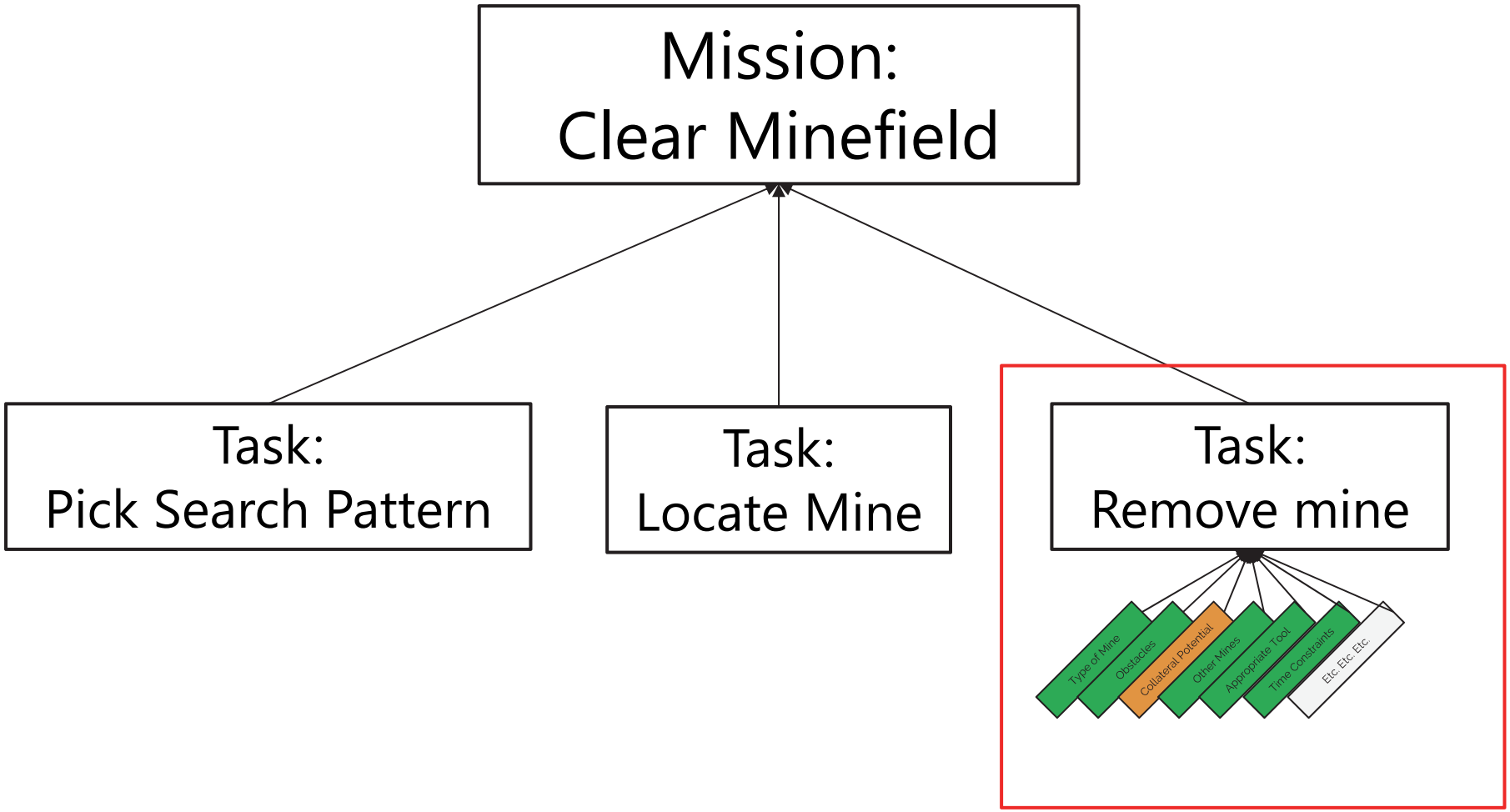
When autonomy is simple, the interpolative, model-based approach will mimic current test designs. When autonomy is complex, it will permit feasible levels of test.

Approved for public release; distribution is unlimited.

Autonomy: Making decisions based on environmental input

VS

Autonomy exists at the level of task being considered



Approved for public release; distribution is unlimited.

Some systems won't need testing to find Goals or Process.

Approved for public release; distribution is unlimited.

Some systems' operationally relevant goals are decided by a human

- Procedural Autonomy
  - ➢ System has autonomy in the moment-to-moment Process decisions to achieve a Goal
    - ❑ E.g., control loops

Approved for public release; distribution is unlimited.

Systems with just Procedural Autonomy don't need special test methods in most cases

- The Process is known in advance
  - Physics model or explicitly coded logic
- Brute force is feasible
  - Small operational space or low-risk consequences
- Correct moment-to-moment decisions just affect performance of a defined task
  - If it is performing well, it is making the right decisions

Approved for public release; distribution is unlimited.

Other systems make decisions about their goals

- Executive Autonomy
  - System can set a goal for itself
  - Making "should" decisions about tasks
- "Should" decision correctness usually won't be captured by typical objective performance metrics
- This is the type of autonomy that really worries people

Approved for public release; distribution is unlimited.

The goal of testing should be developing and confirming a generalizable model of system decision-making.

- Brute force testing is not feasible
- Interpolation required for evaluation and TTPs
- Models enable interpolation
- Solution: structure tests and pick test points based on what allows you to understand the decision process

Approved for public release; distribution is unlimited.

The lifecycle of test must be a continuum, not discrete categories operating as independent fiefdoms.

Testing autonomy will require more data

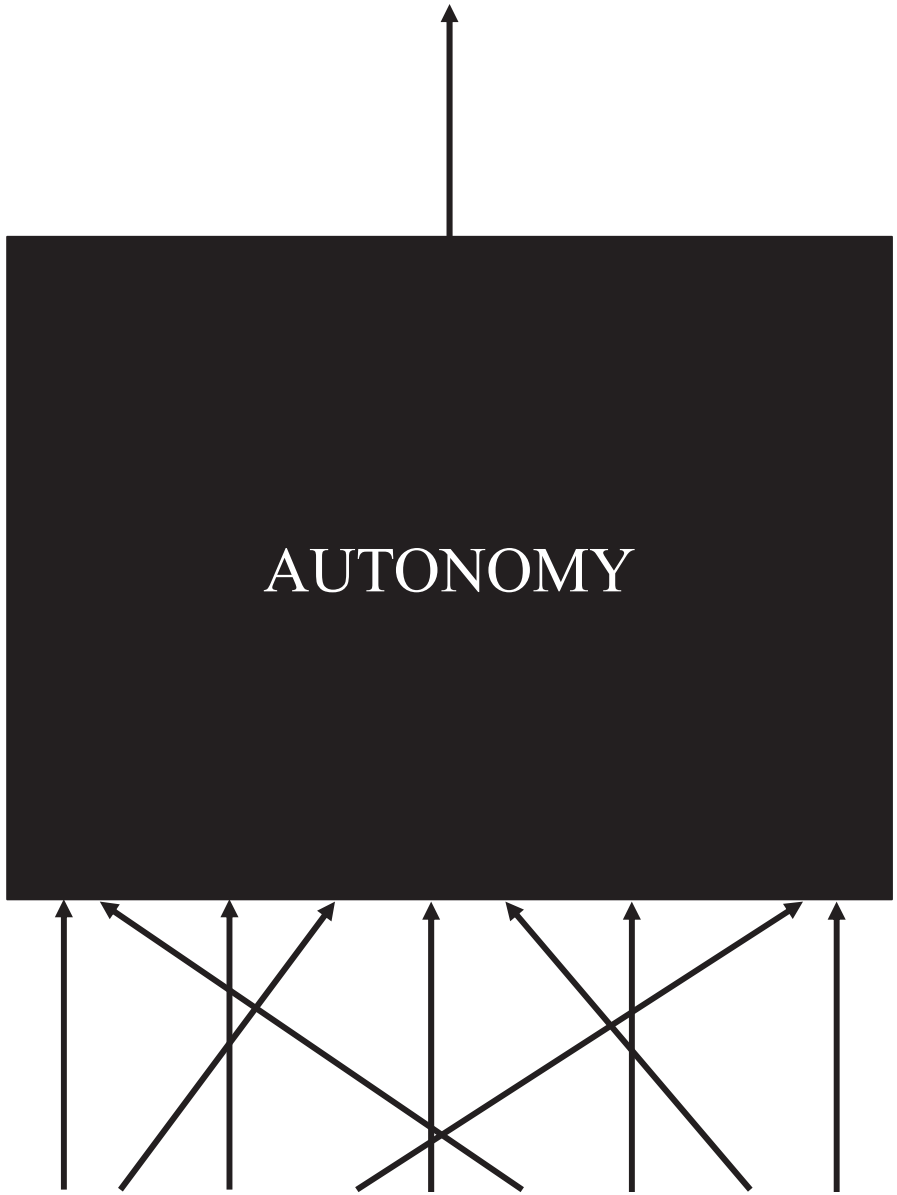
- Autonomous systems will have larger operational spaces
  - Still have to test physical performance
  - Also have to test decision performance
    - Adds (many) factors to test design
- People likely less forgiving of machine decisions
  - Acceptable level of risk will be smaller
    - Requires more evidence to achieve acceptable risk
- Need efficient methods to discover AI's model

Approved for public release; distribution is unlimited.

Designing a model is easier than figuring it out.

Approved for public release; distribution is unlimited.

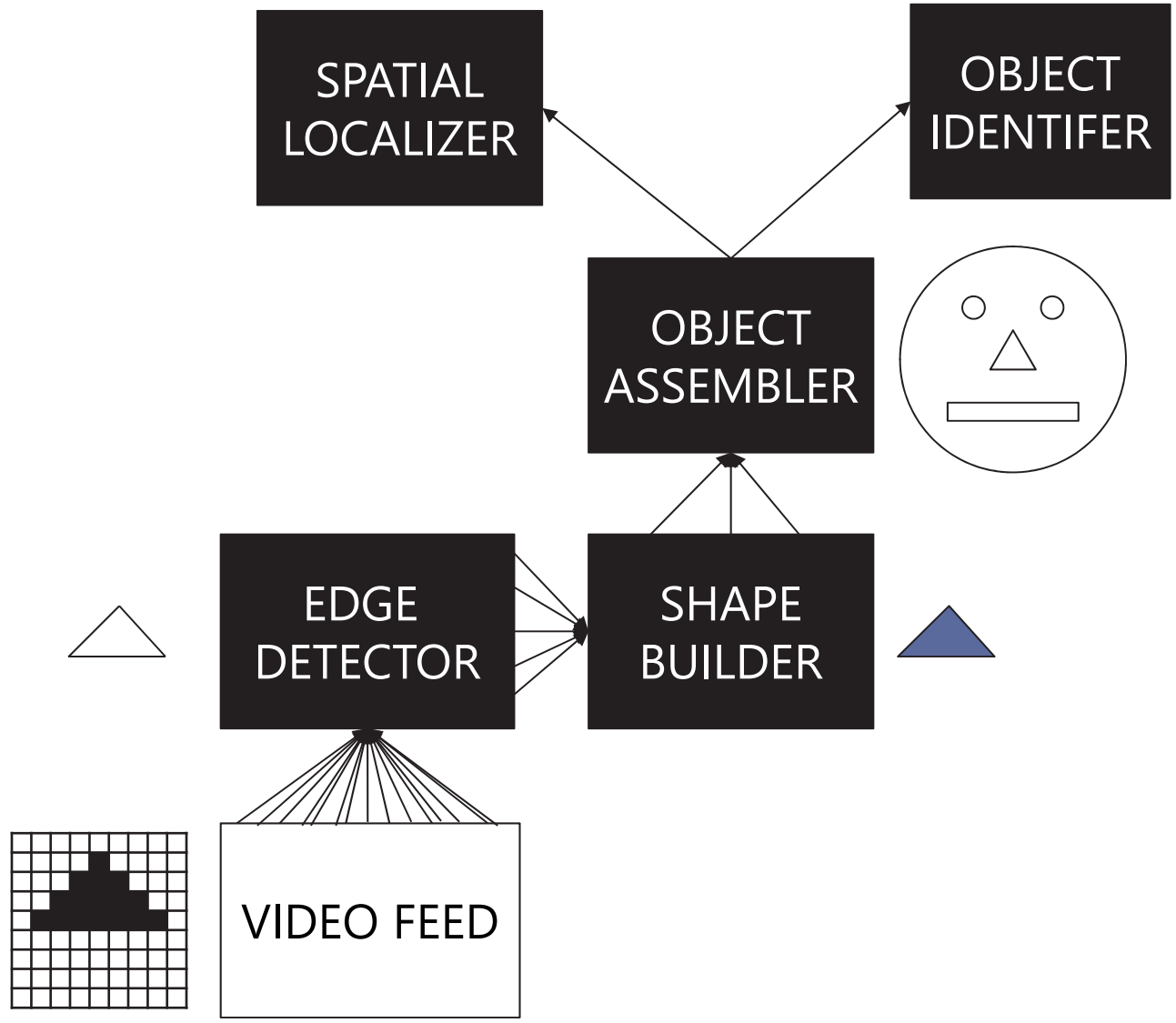
The Black (Box) Plague should be avoided



Approved for public release; distribution is unlimited.

Cognitive Architecture: Life's easier if you plan!

Task: Identify what things are where



We must get more evidence without breaking budgets

- **Challenge:** Autonomy will require more evidence
  - More evidence doesn't have to mean more formal test points
- **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 system
  - Expand data sources that inform operational evaluation
    - More evidence without more test

Approved for public release; distribution is unlimited.

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

Approved for public release; distribution is unlimited.

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

Approved for public release; distribution is unlimited.

DT-OT should be a continuum

Figure out what we think the model is

Try to disprove model under realistic conditions

Approved for public release; distribution is unlimited.

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

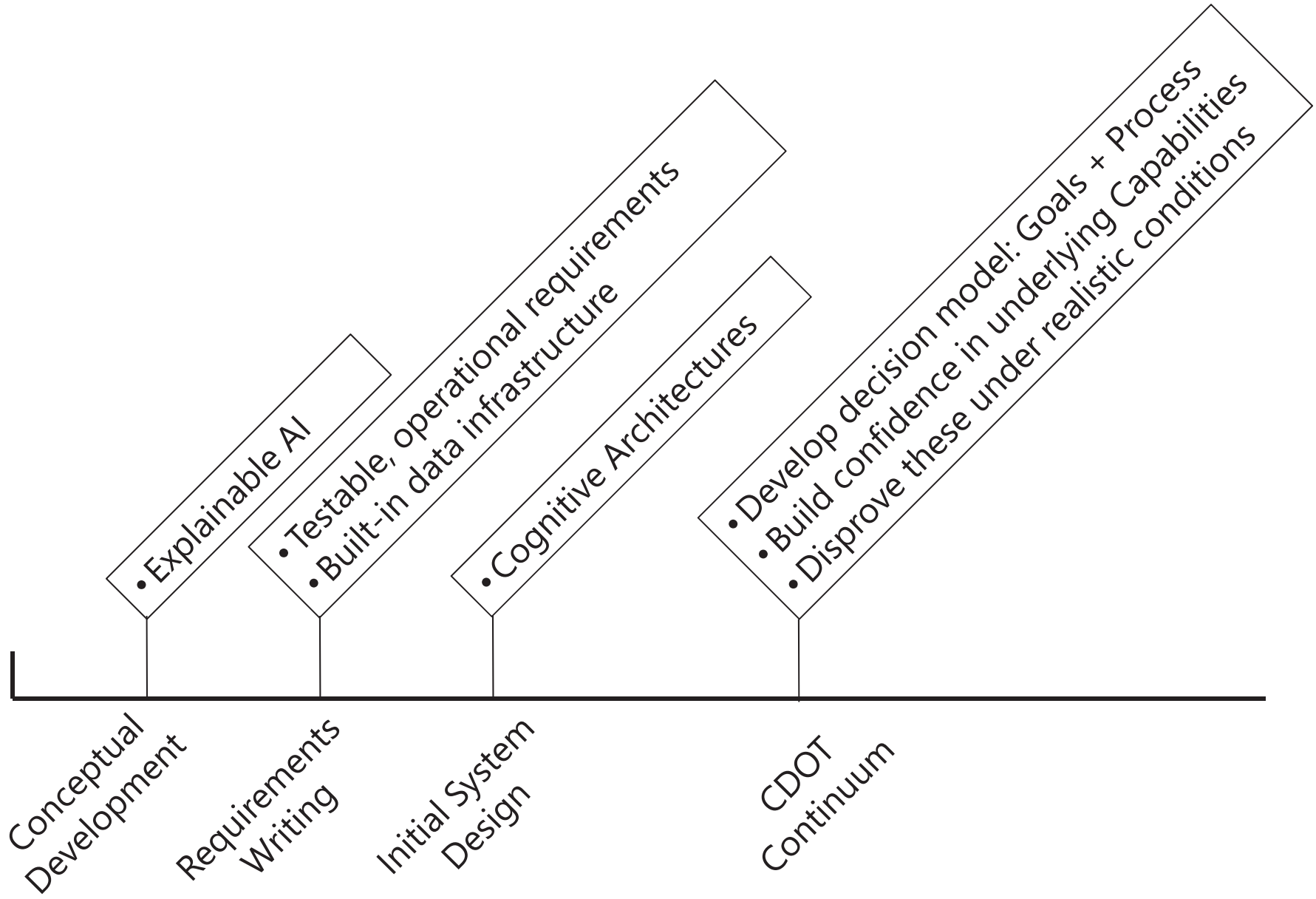
- **Solution:** Limited or Incremental Capability Fielding
  - Complex tasks can be broken down into smaller ones
  - Choose a subtask with acceptable risk and test that
    - If it passes this test, approve it for fielding on that task
      - Potentially limit to human supervision
  - Collect field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

Approved for public release; distribution is unlimited.

Data collection must be built into 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

Approved for public release; distribution is unlimited.



Approved for public release; distribution is unlimited.

Talk Takeaways – Order of Importance

1. The goal of testing should be developing and confirming a generalizable model of system decision-making.

2. The lifecycle of test must be a continuum, not discrete categories operating as independent fiefdoms.

3. The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations.

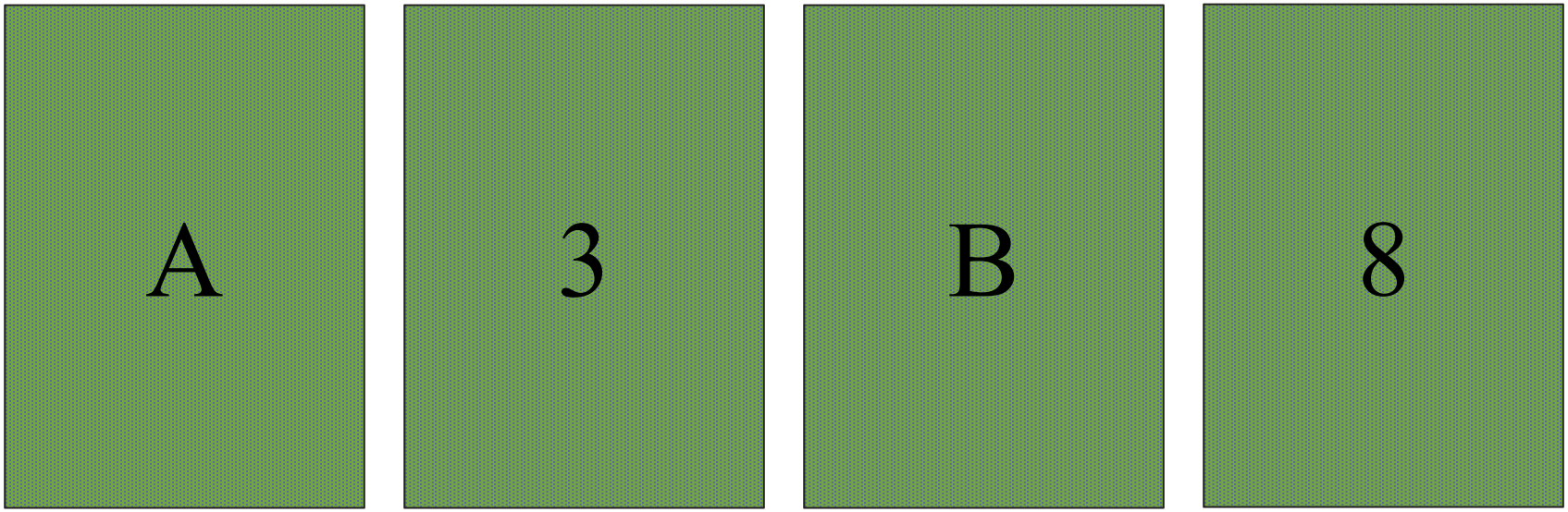
Approved for public release; distribution is unlimited.

Thank You

Approved for public release; distribution is unlimited.

Backup

All *vowels* have an *odd* number on the back. What cards do you flip to test this most efficiently?



Approved for public release; distribution is unlimited.

The fundamental challenge of testing autonomy and AI is generalizing to unobserved situations

Situation #1
Strangers
Bidirectional Blvd
Shadows & Night
Pickup Friend
Streetlights
People

Situation #2
Familiar Faces
Worn-out Road
Bright & Early
Daily Commute
Stuck in Traffic
People

Situation #N
Strangers
Two-lane Road
Shadows
Night
No Streetlights
People

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
27-03-2019	OED Draft	

4. TITLE AND SUBTITLE

Demystifying the Black Box: A Test Strategy for Autonomy

5a. CONTRACT NUMBER

HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Daniel J. Porter

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-10465-NS

H 2019-000126

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 are beginning to permeate civilian, industrial, and military sectors. Though these technologies have the potential to revolutionize our world, they also bring a host of new challenges in evaluating whether these tools are safe, effective, and reliable. The Institute for Defense Analyses is developing methodologies to enable testing systems that can, to some extent, think for themselves. In this talk, we share how we think about this problem and how this framing can help you develop a test strategy for your own domain.

15. SUBJECT TERMS

Autonomy; AI; DATAWorks 2019; Artificial Intelligence; 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	49	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6811

Approved for public release; distribution is unlimited.