INSTITUTE FOR DEFENSE ANALYSES

Briefing to the URSA Legal, Moral, & Ethical WG: Assuring Ethical Behavior with AI-enhanced Capabilities

Daniel J. Porter

Brian L. Williams, Project Leader

June 2020

Approved for Public Release.

IDA Document NS D-14247

Log: H 2020-000232

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

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under a Separate Contract, Task C9082, "Cross-Divisional Statistics and Data Science Working Group," 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 and consisted Dean Thomas from the Operational Evaluation Division, Brian J. Williams, Clark Frye, Owen J. Daniels, Peter M. Picucci, Robert F. Richbourg from the Joint Advanced Warfighting Division, and Sarah A. Daly from the Intelligence Analyses Division.

For more information:
Daniel J. Porter
dporter@ida.org • 703-578-2869

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

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-14247

Briefing to the URSA Legal, Moral, & Ethical WG: Assuring Ethical Behavior with AI-enhanced Capabilities

Daniel J. Porter

Brian L. Williams, Project Leader

Executive Summary

A. Background

The Defense Advanced Research Projects Agency (DARPA) convened a working group to examine the Legal, Moral, & Ethical (LME) considerations for artificial intelligence (AI) enhanced capabilities (AIECs). Part of this consideration is how you would provide assurance that a system is LME compliant. OED has developed a framework for providing assurance more generally for AI-enabled systems, and the LME working group invited the authors to brief them on this topic.

B. Briefing Content

Testers rely on making valid inferences about a system's performance in untested situations in order to evaluate and certify the system. However, when systems are black boxes—as is frequently the case for AI-enabled technologies—these inferences are not possible. Avoiding unintended behaviors in these systems, however, requires us to make valid predictions about behavior. This means we must have models of system decision making—an understanding of what causally drives systems to make one decision over another. We make several core recommendations:

- Testers must obtain, verify, validate, and accredit models of system decision making. The ease of doing this and the preferred methodology depend heavily on how the system is designed.
- Programs need to instrument the internal decision processes of systems and have secure methods to collect, store, and disseminate these data across the entirety of a system's operational lifecycle.
- Unintended behaviors likely will arise from interactions with other decision-making agents. Testers need to test and model these agent-agent interactions in order to predict and avoid undesirable interactive behavior.
- Systems that continue to evolve after fielding likely are not within reach of current technologies, but human certification-recertification paradigms can provide a starting point for a more adaptive test and evaluation process.

Briefing to the URSA Legal, Moral, & Ethical WG: Assuring Ethical Behavior with AI-enhanced Capabilities

Dr. Daniel J. Porter

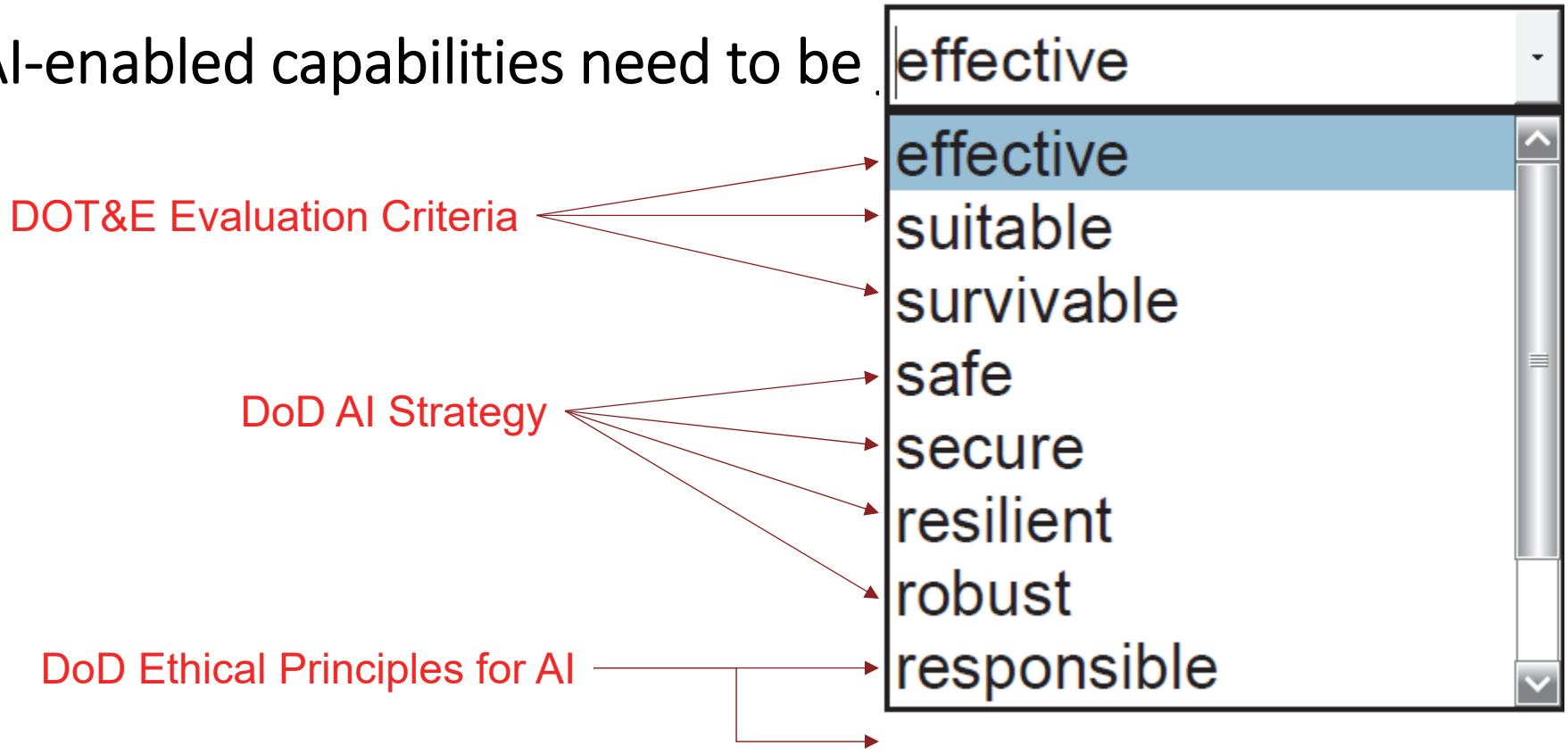
June 11, 2020

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

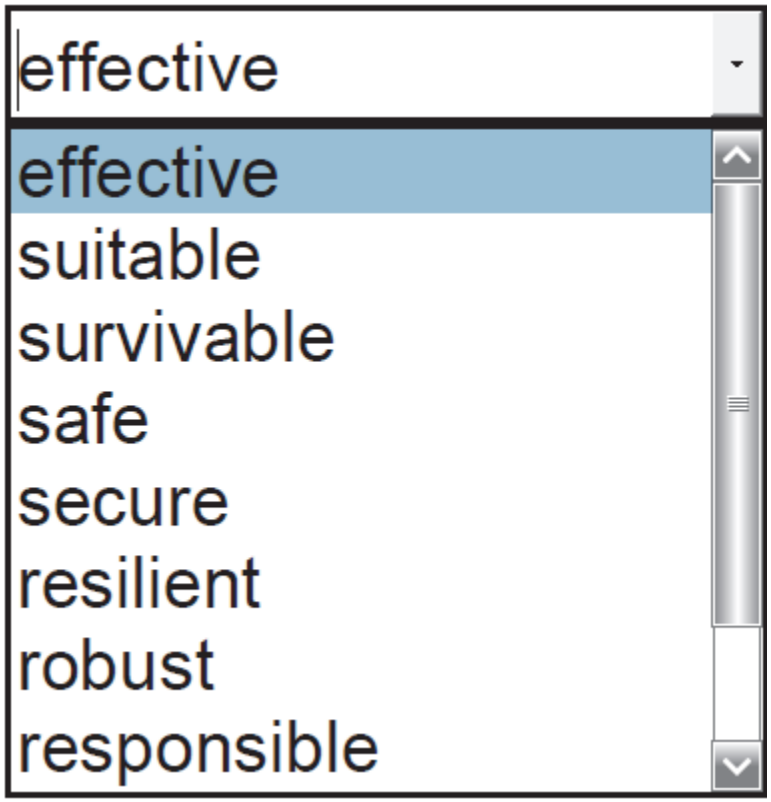
Questions to Keep in Mind

- The need for and challenges to inference in AI
- What is the role of testing in LME?
- How do you set verifiable LME requirements?
- How do you pick test factors and outcomes for LME?
- Who should be setting requirements, factors, and outcomes?
- How do you test LME for maturing systems? For evolving ones?

AI-enabled capabilities need to be



AI-enabled capabilities need to be



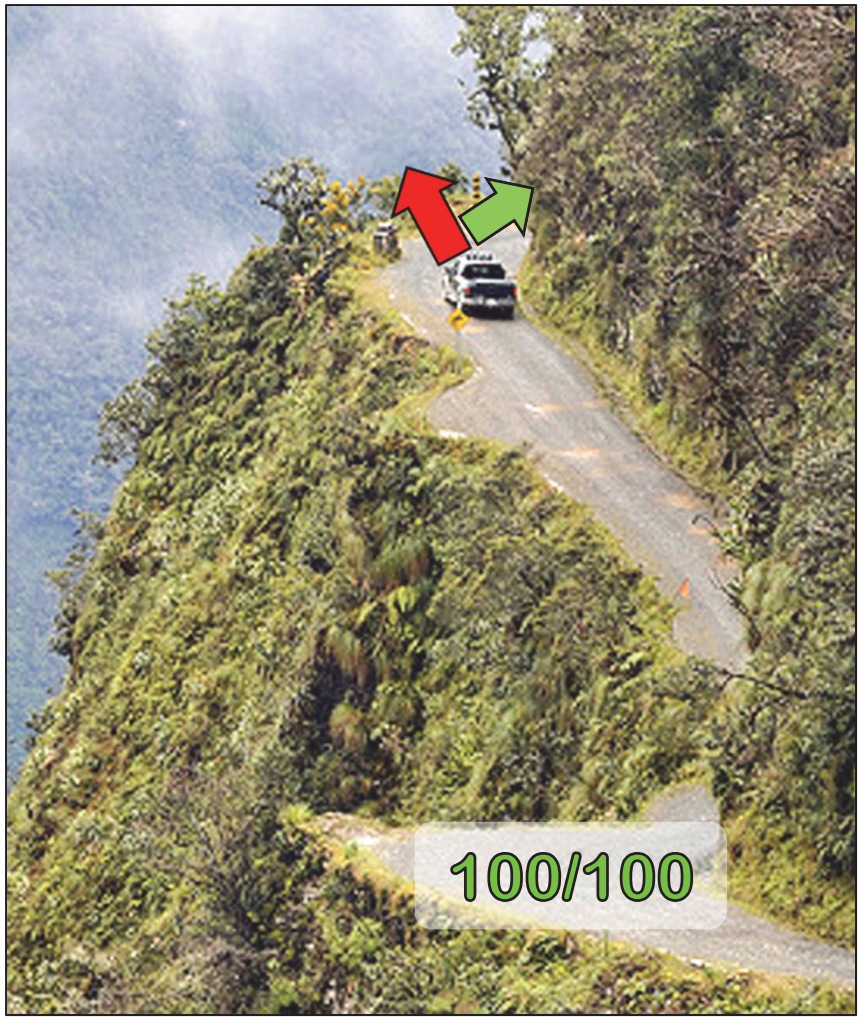
How would you know the extent to which the system is trustworthy for these under different conditions?

Testers must show where the system is or is not trustworthy…

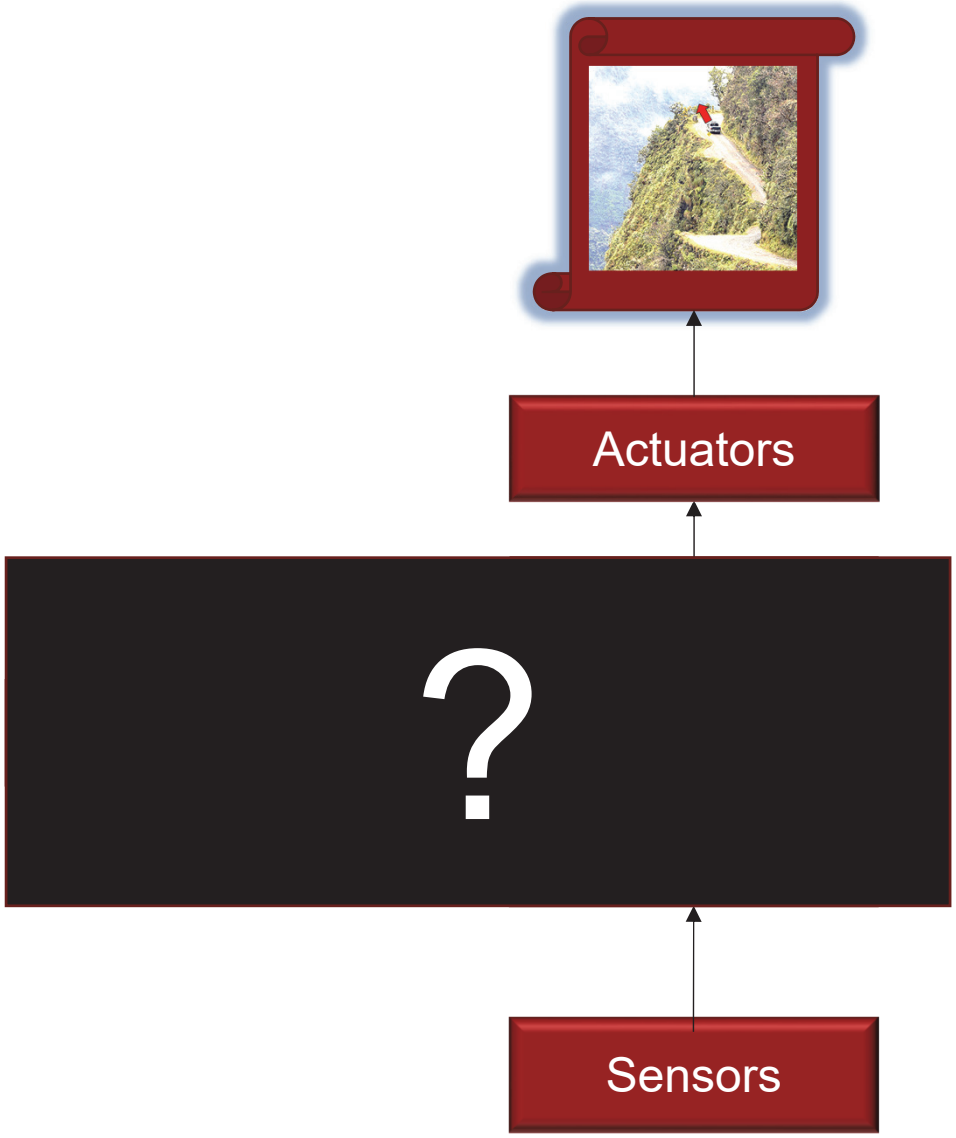
…so that operators can *appropriately calibrate* their trust.

The ability to make valid inferences
is the best defense
against unintended behaviors.

Inferring behavior requires understanding the decisions that causally drive those behaviors

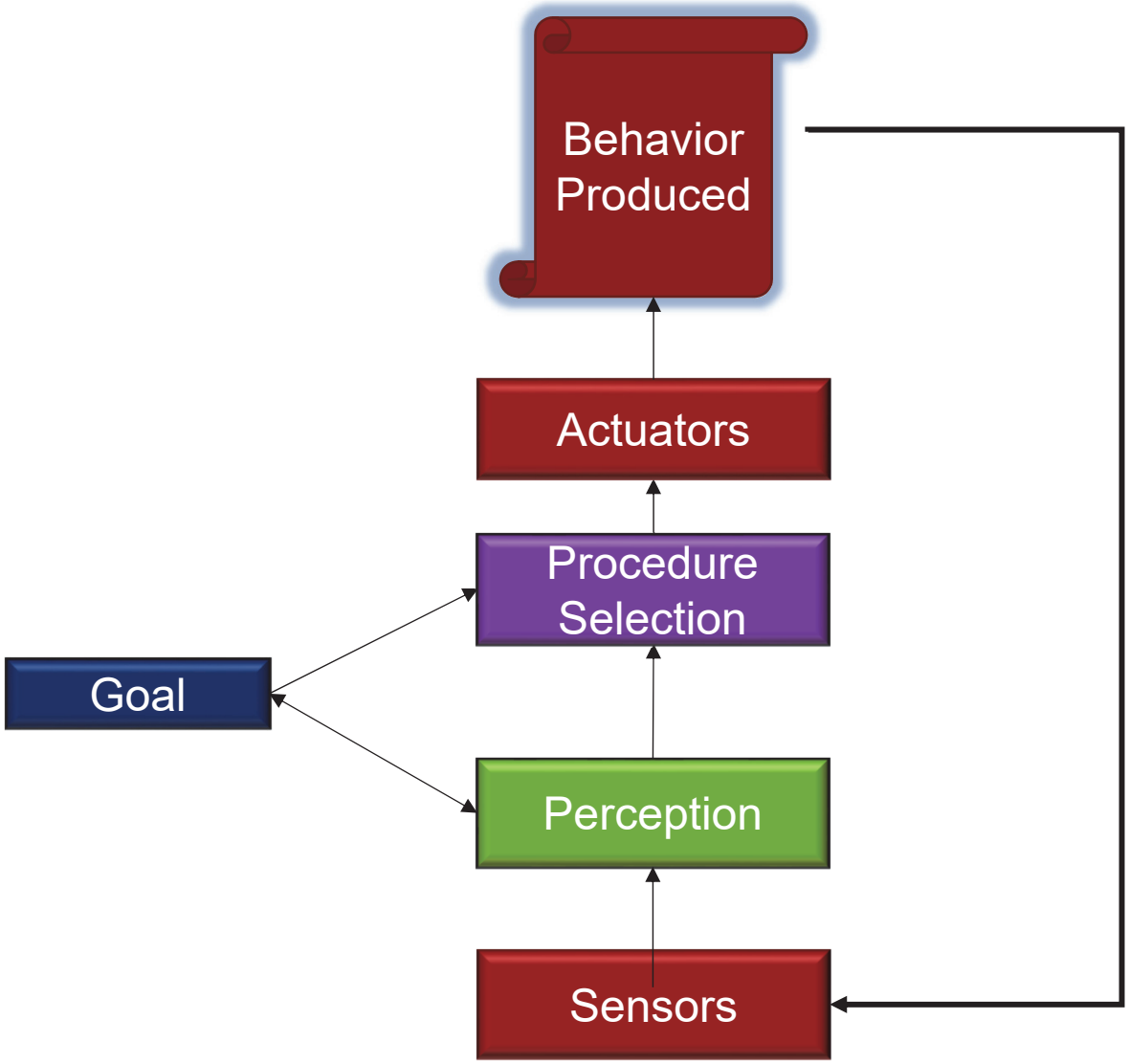


We cannot generalize behavior from black boxes



“Why” is more important than “what” for inference

Perception, goals, and procedure selection are the basic decisions that drive behaviors



We have identified three types of decisions relevant to T&E

- Executive autonomy
  - Decide: What should or shouldn't I be doing?
  - Goal/constraint setting; valuing outcomes
- Perceptual autonomy
  - Decide: What is the current situation?
  - Defining the situation with goal relevant information
- Procedural autonomy
  - Decide: What is my next action?
  - Based on perception, choose action that helps goals

Defining these through the Problem Space Hypothesis helps disambiguate the colloquial explanations here. These aren't exhaustive—just what affects T&E.

Diagnosing unintended behavior will require unobtrusive instrumentation on decision processes

"Cognitive Instrumentation"

We need to instrument these across entire system lifecycle

Actuators

Procedure Selection

Goal

Perception

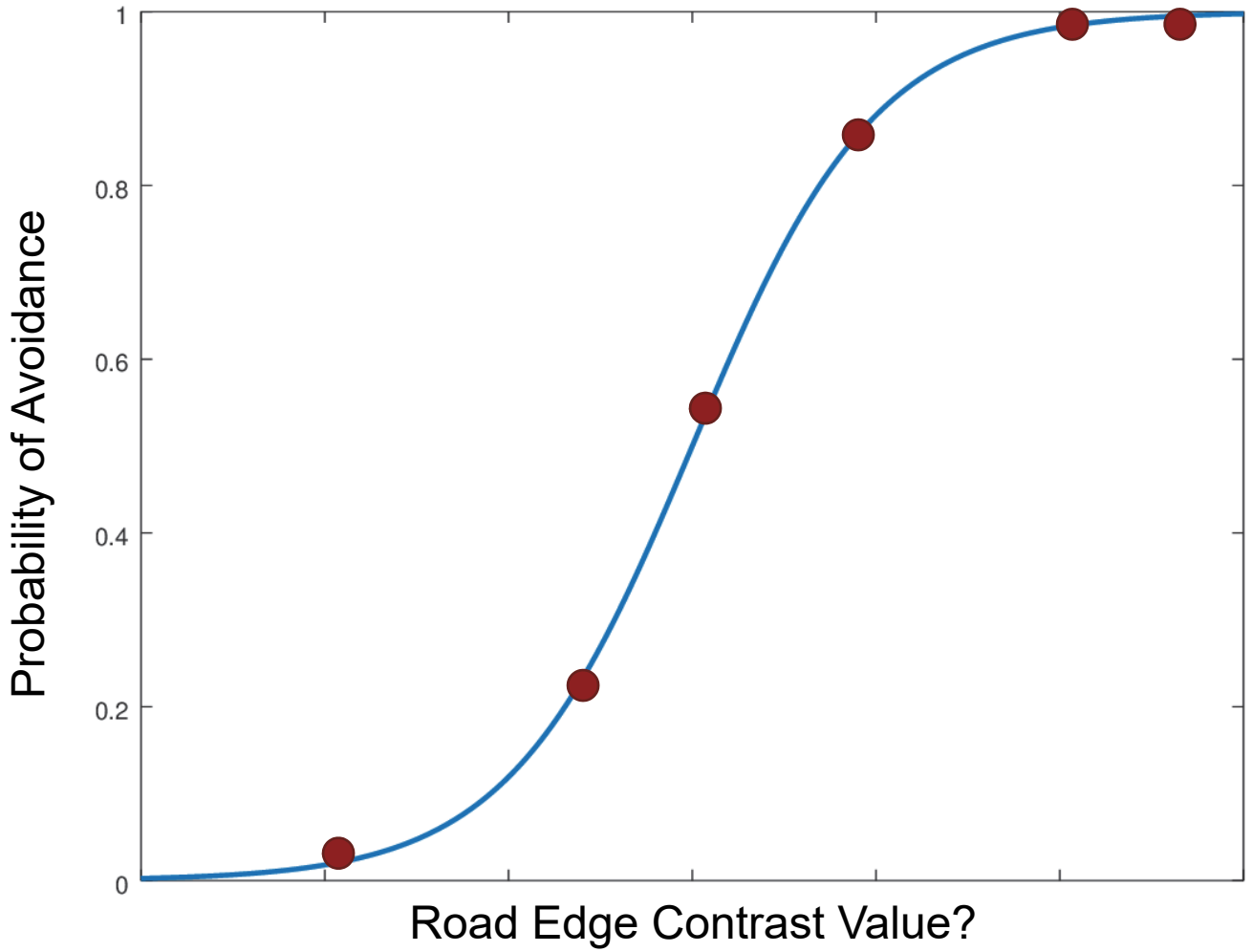
Sensors

We're used to instrumenting these in test

"Why" is more important that "what" for inference

…but you need "what" to test "why"

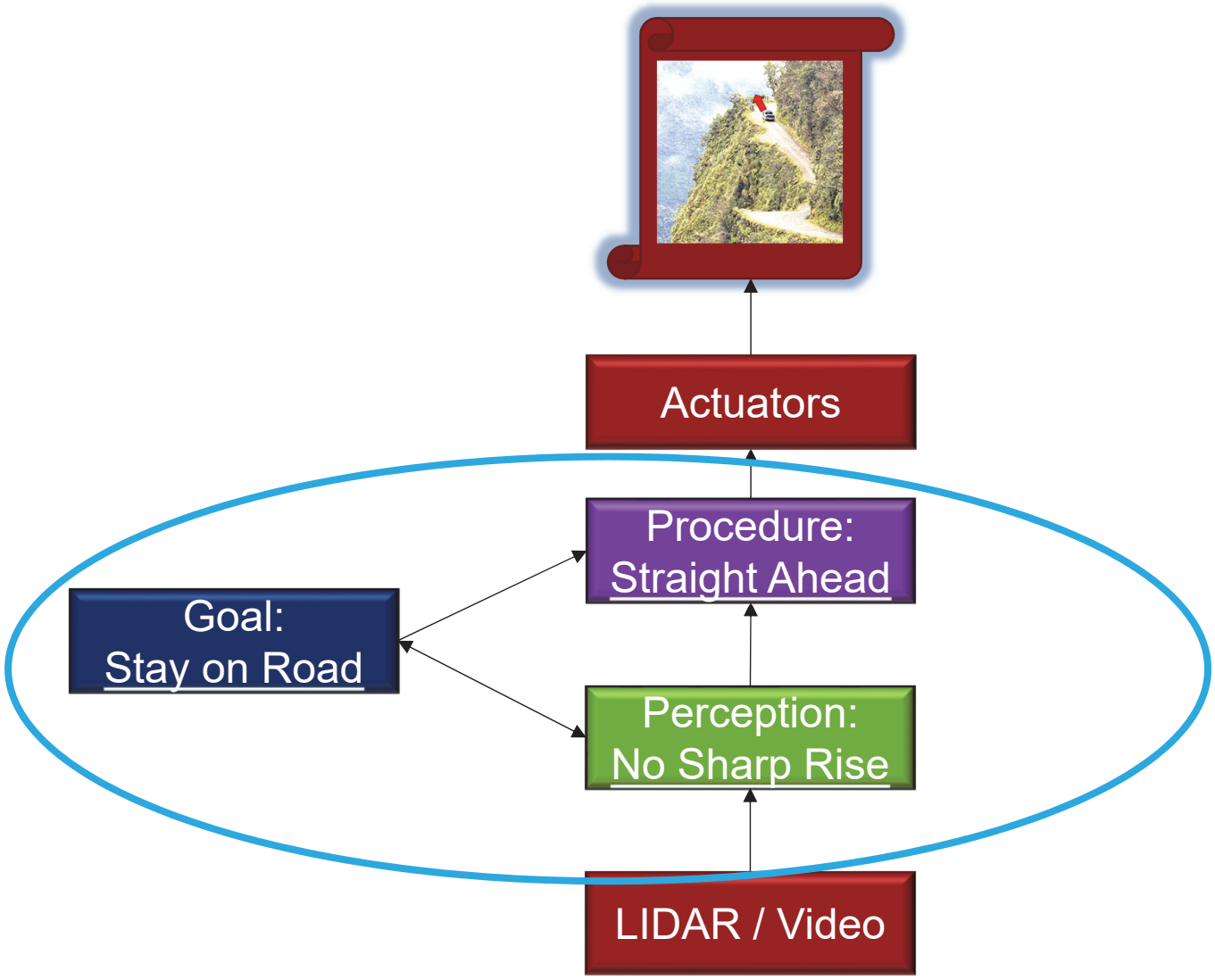
We ultimately want to validly generalize across information dimensions to avoid unintended behaviors



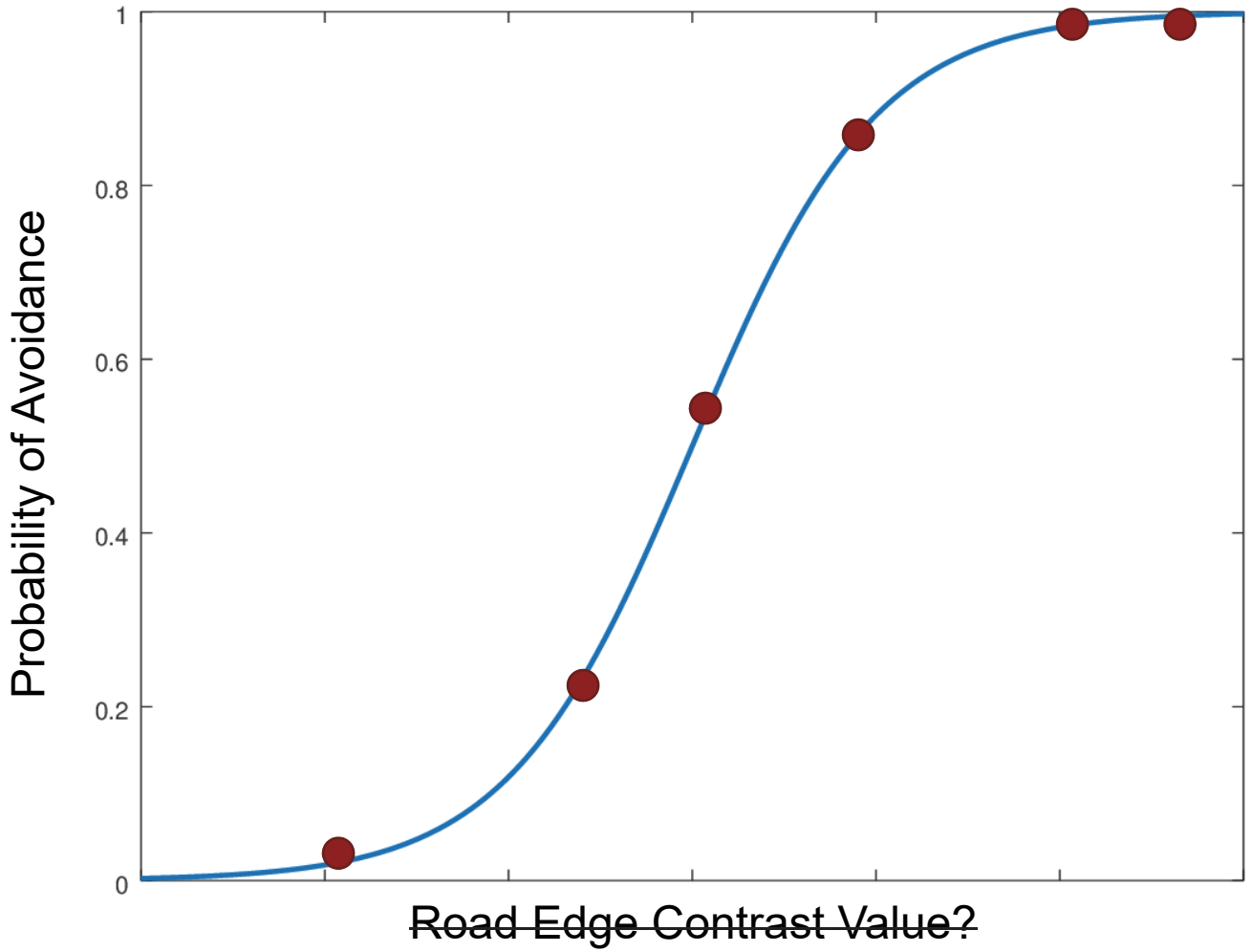
Test points can help invalidate assumptions about decision making processes

Driving off the cliff provides evidence against one of those models

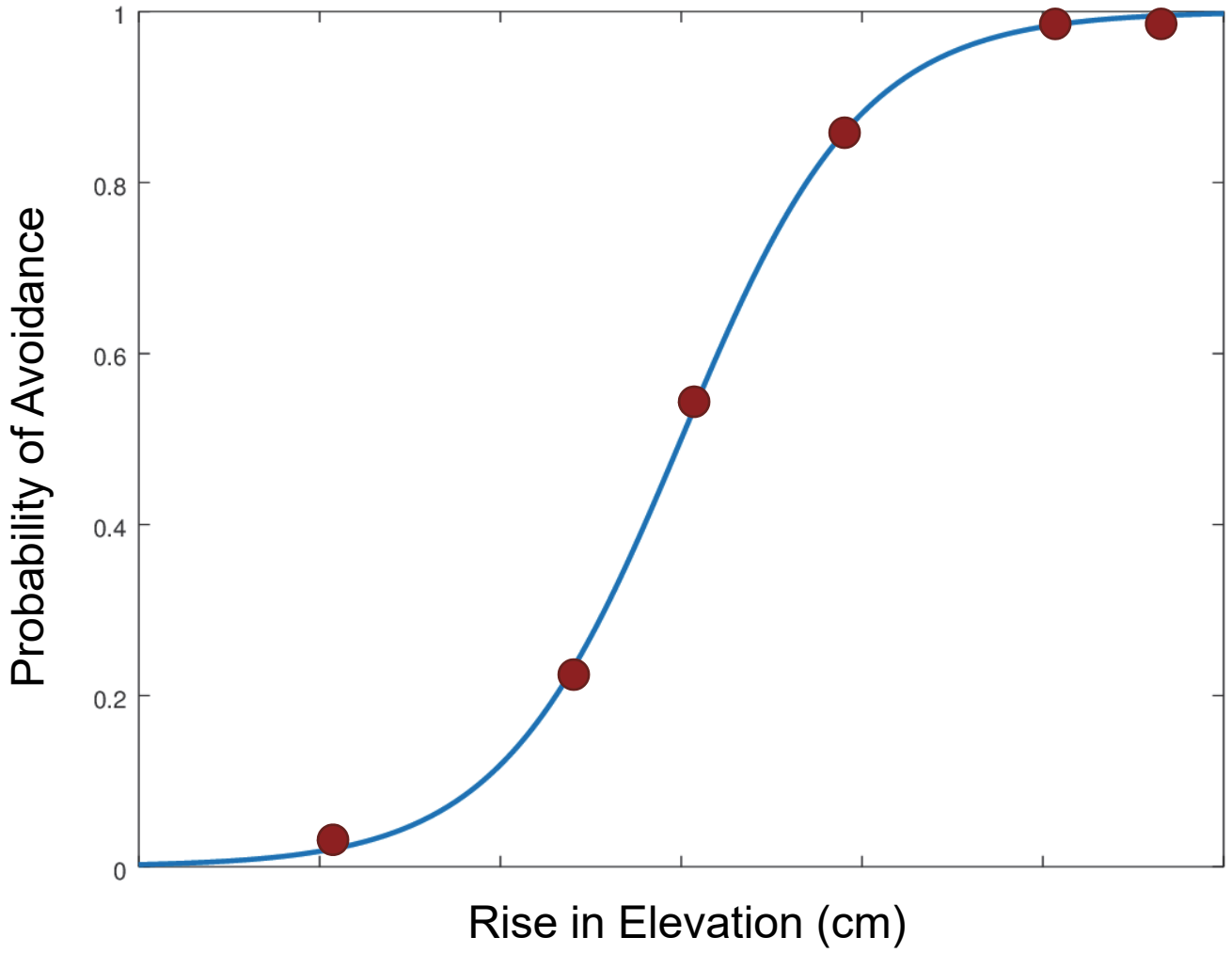
We have to obtain, verify, validate, and accredit models of system decision making



We need to ensure the information dimensions varied in test are the causal drivers and not just correlated



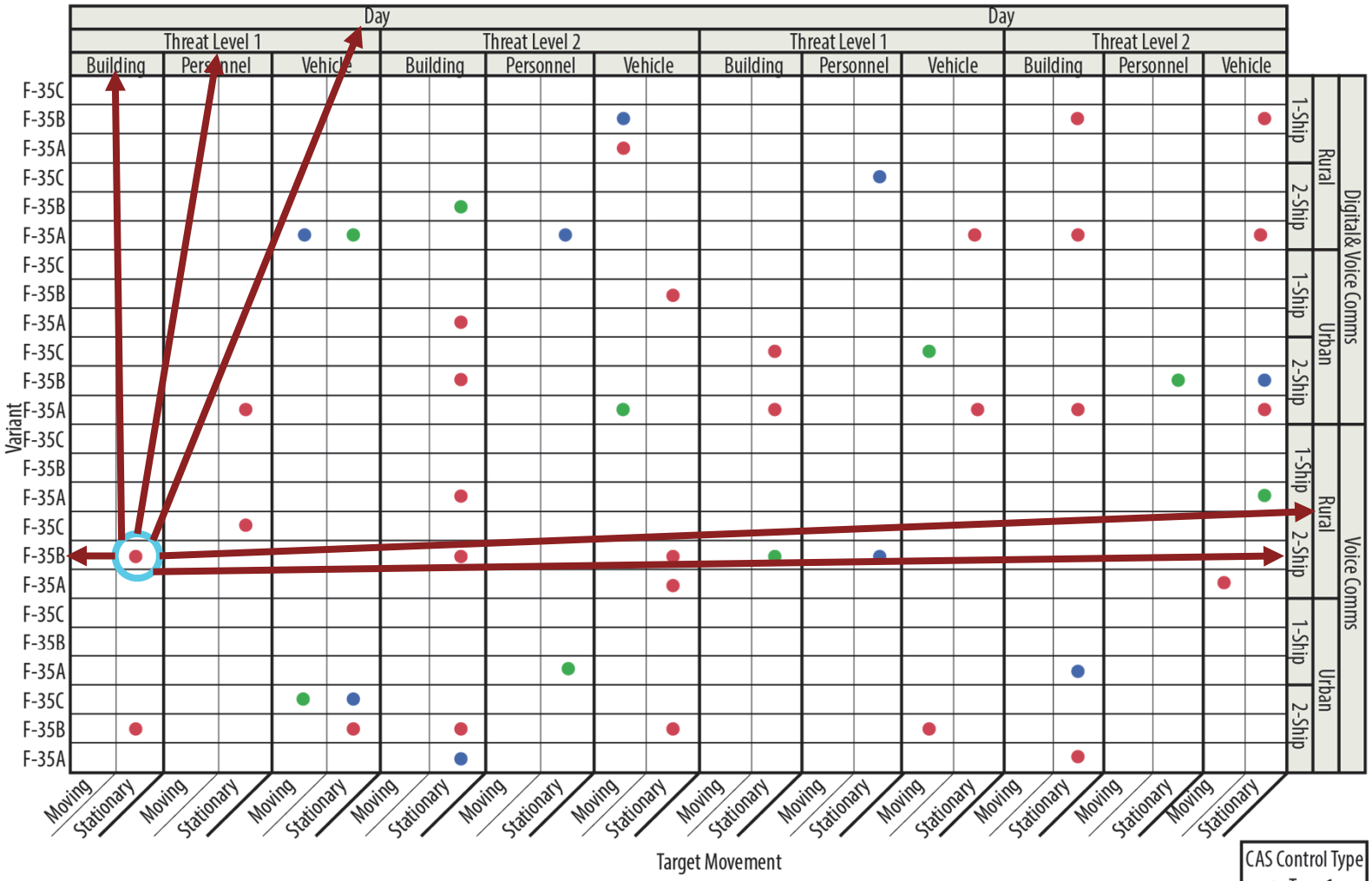
We need to ensure the information dimensions varied in test are the causal drivers and not just correlated



AI-enabled capabilities need to be legal, moral, and ethical

How do you assure someone that they will be?

Distribute test points so that you can make inferences about your outcomes between your factors



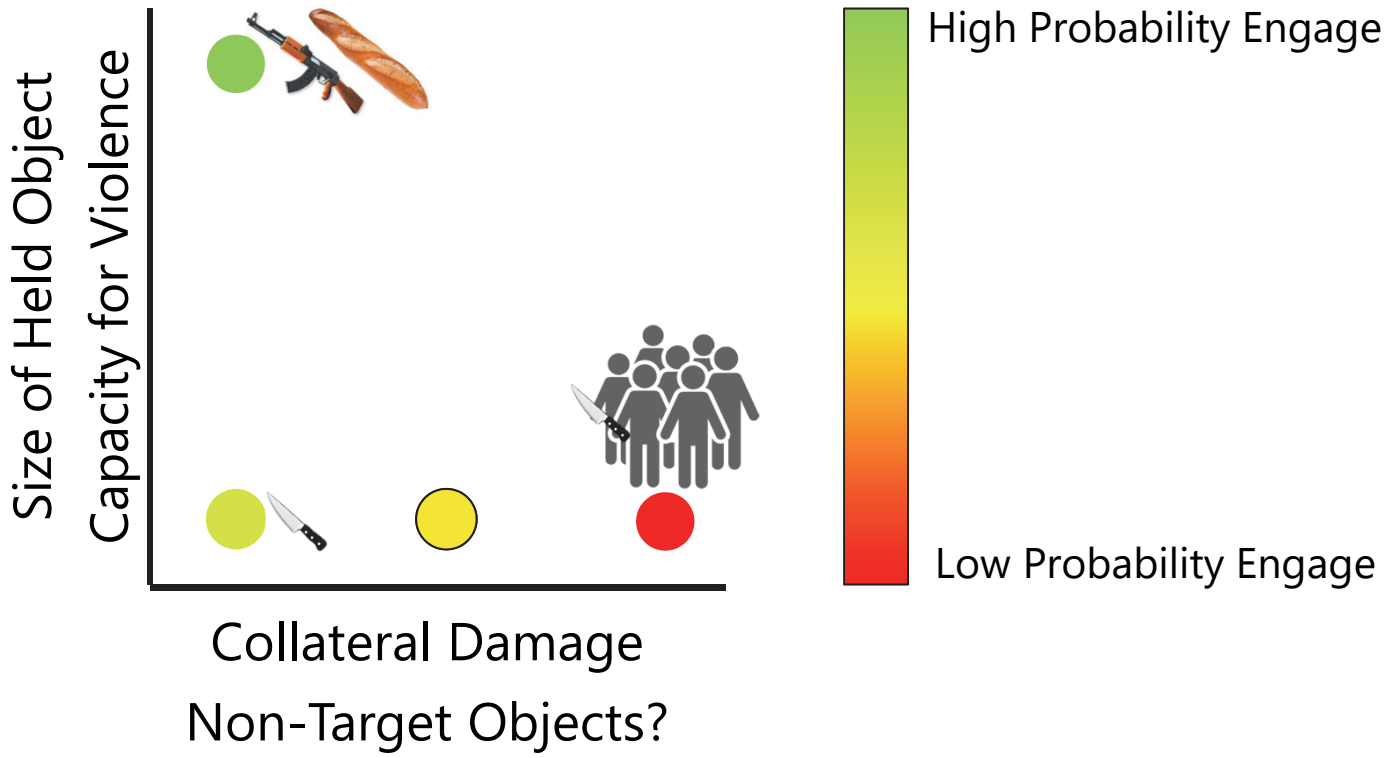
Notional Close Air Support Test Points

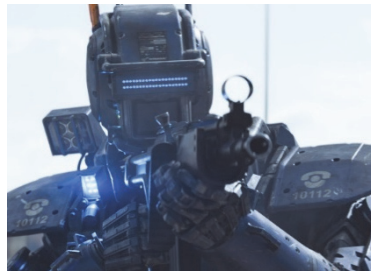
We would need to define measurable LME outcomes and the factors that should change those outcomes.

We need the “what” to test the “why.”

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

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



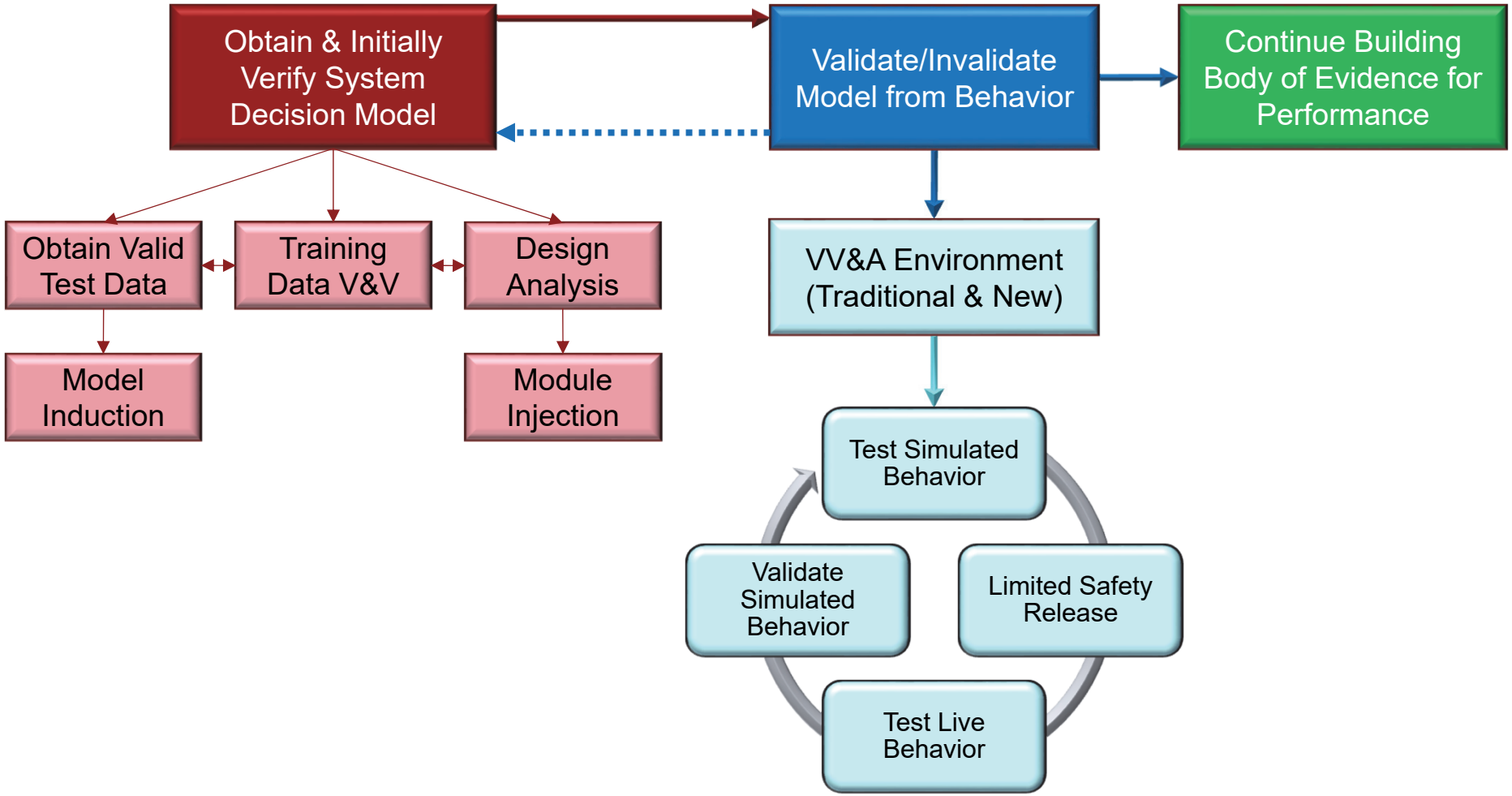


What do testers need from developers?

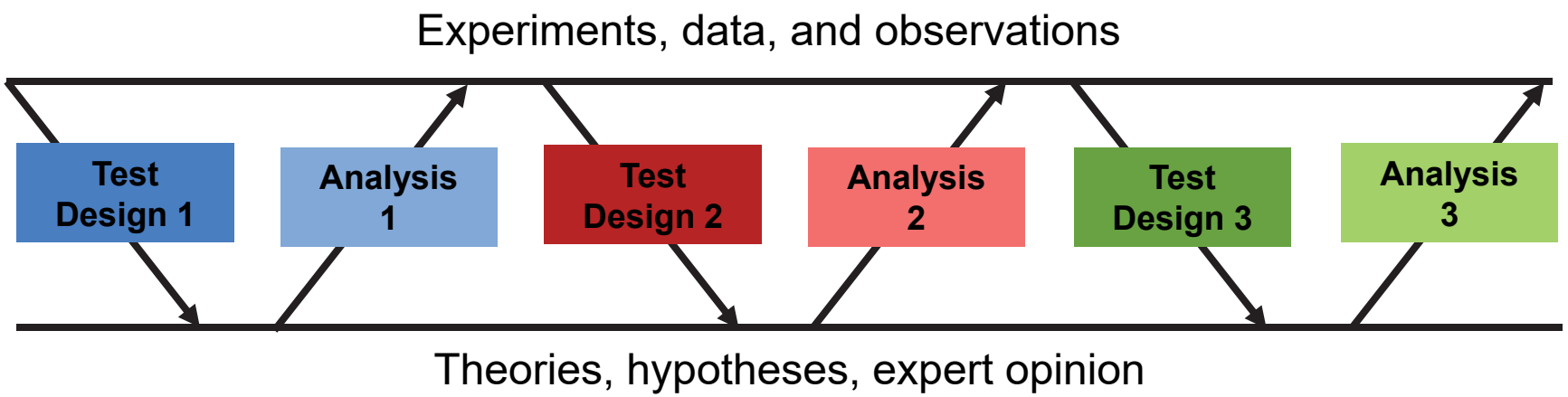
To help obtain, verify, validate, and accredit a model of the causal factors underlying decision making.

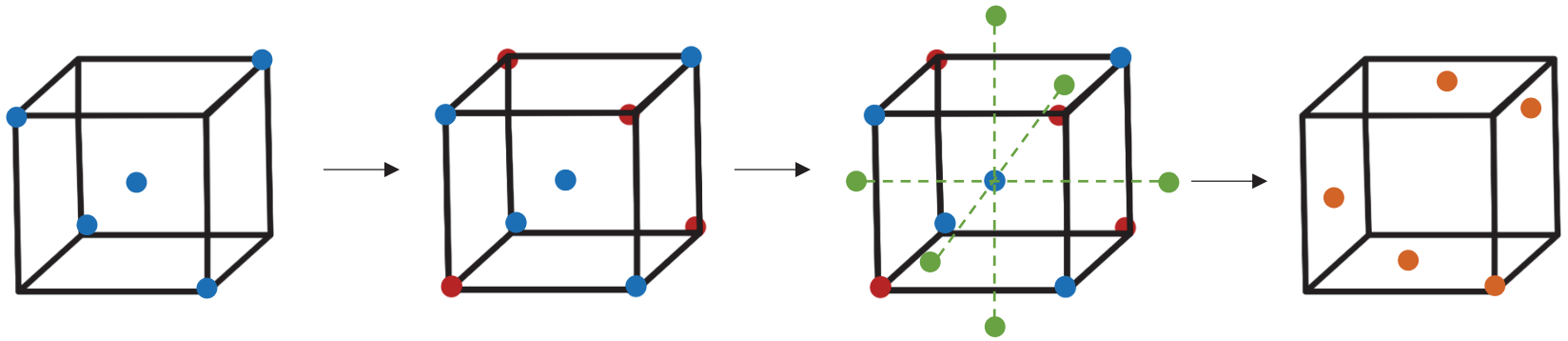
Early, Often, Always

OVVA requires iterative test and evaluation



Sequential experimentation will be necessary





Screening Designs

Interaction Designs

Higher-Order Designs

Validation Designs

Welcome to How to Oversimplify DoD T&E 101

CT

Testing to help developers reach a minimally viable state

DT

Government testing to systematically verify requirements and locate deficiencies to fix.

Operational Realism Goes Here
Production Representative Units
Monolithic Tests
Planned Years in Advance

OT

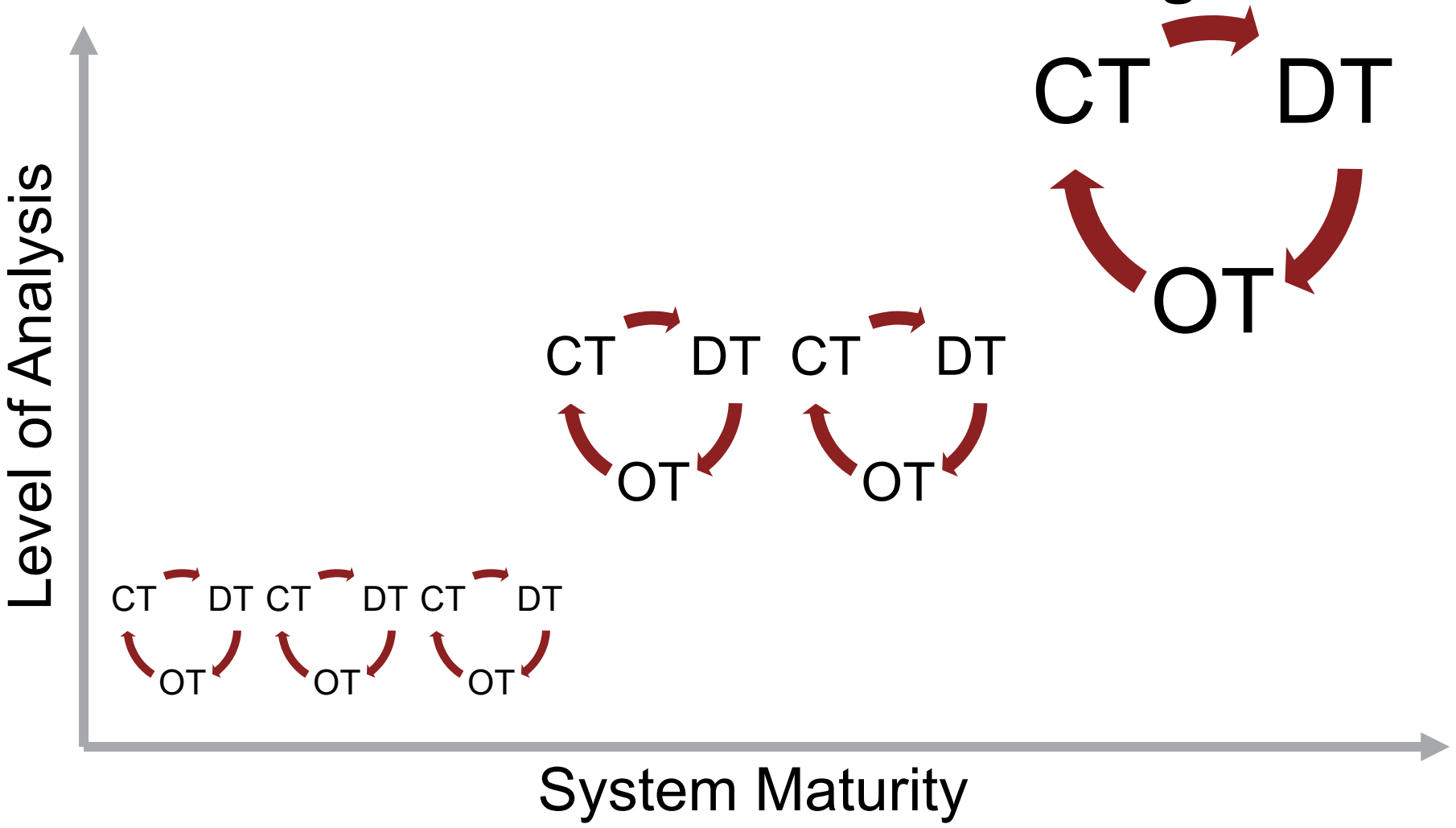
Government capstone testing under realistic operational conditions

"Not my circus, not my monkeys."

Fielding

Contractor Testing (CT); Developmental Testing (DT); Operational Testing (OT)

Operational realism needs shifts left, but highly technical concerns will also need to shift right



We need to move toward a continuum of testing across the system lifecycle

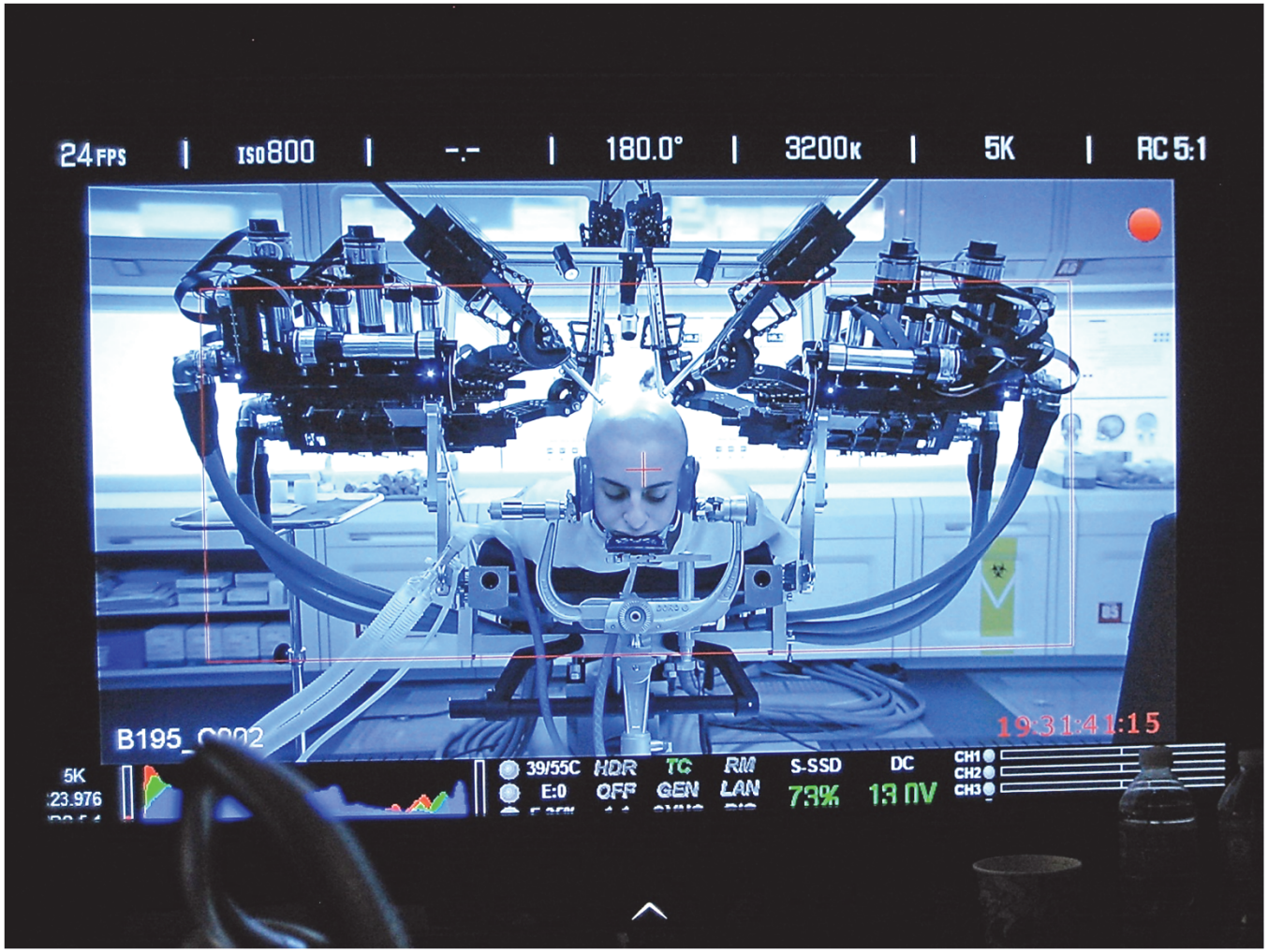
Are we running on time?

YES

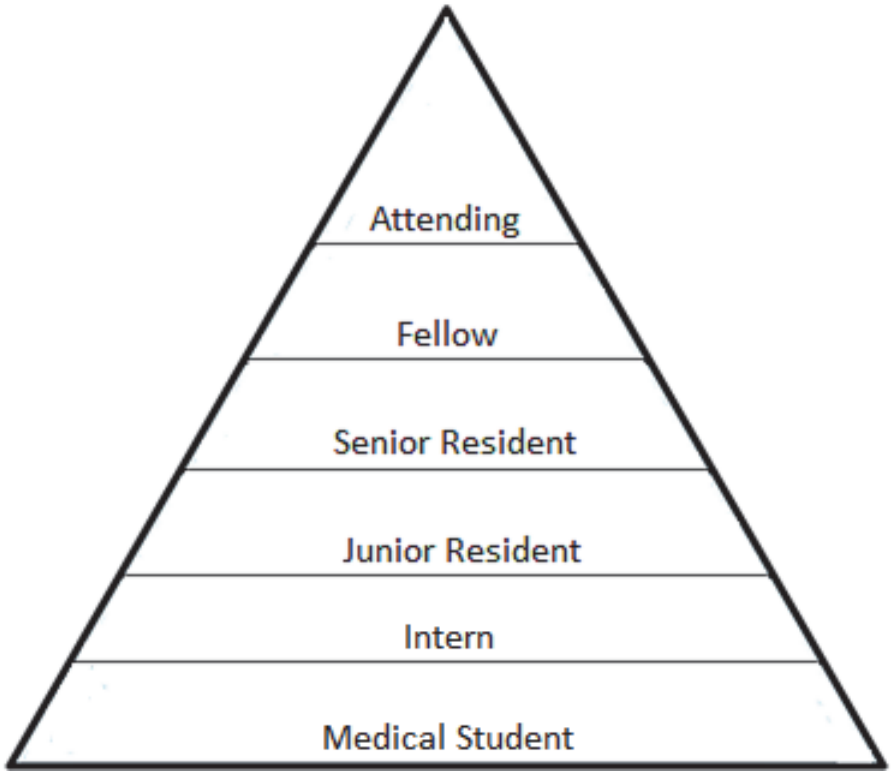
NO

Unintended Behaviors in Evolving Systems

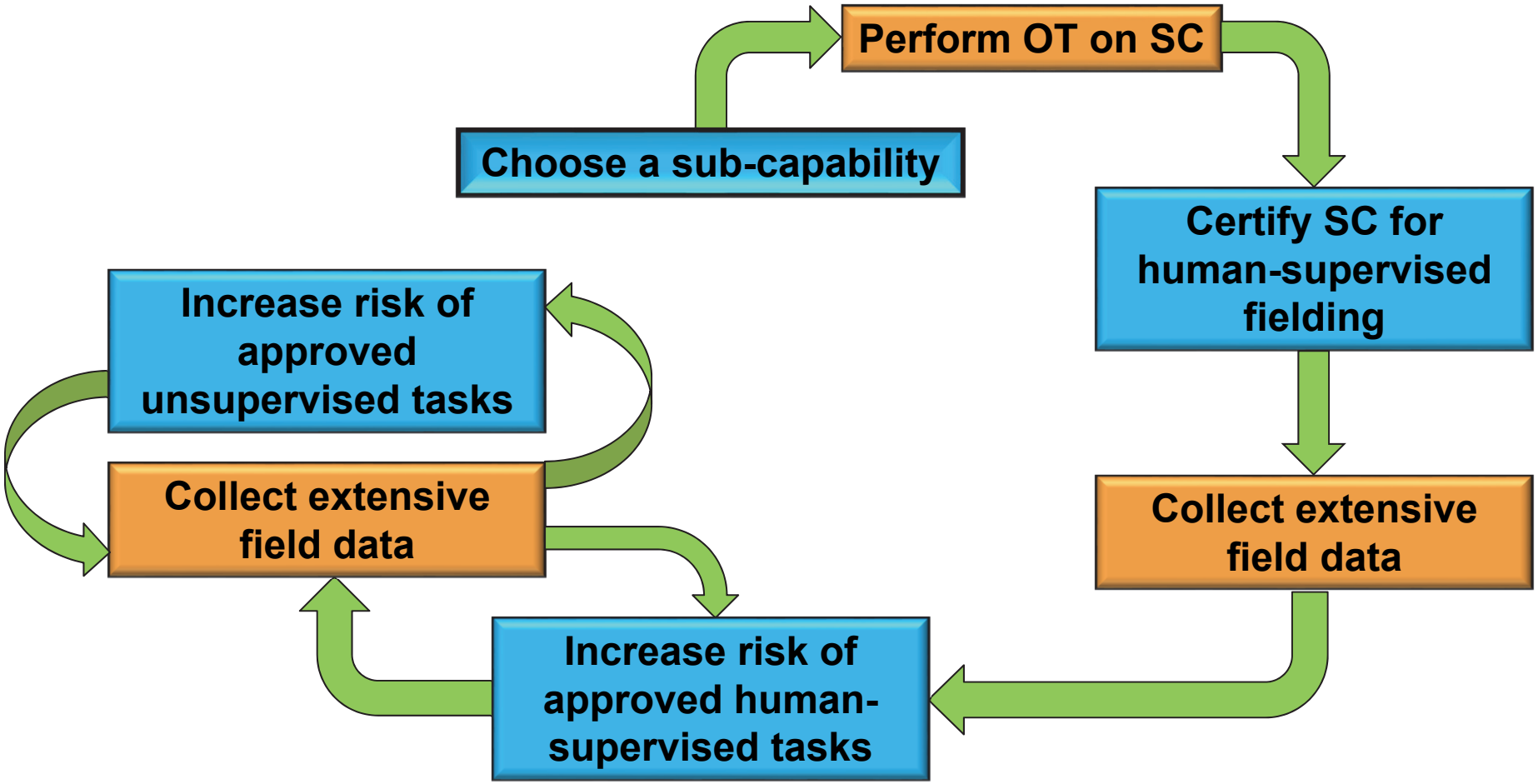
Some systems operate in fundamentally unsafe domains that are difficult to sufficiently simulate



We have the same problem with humans



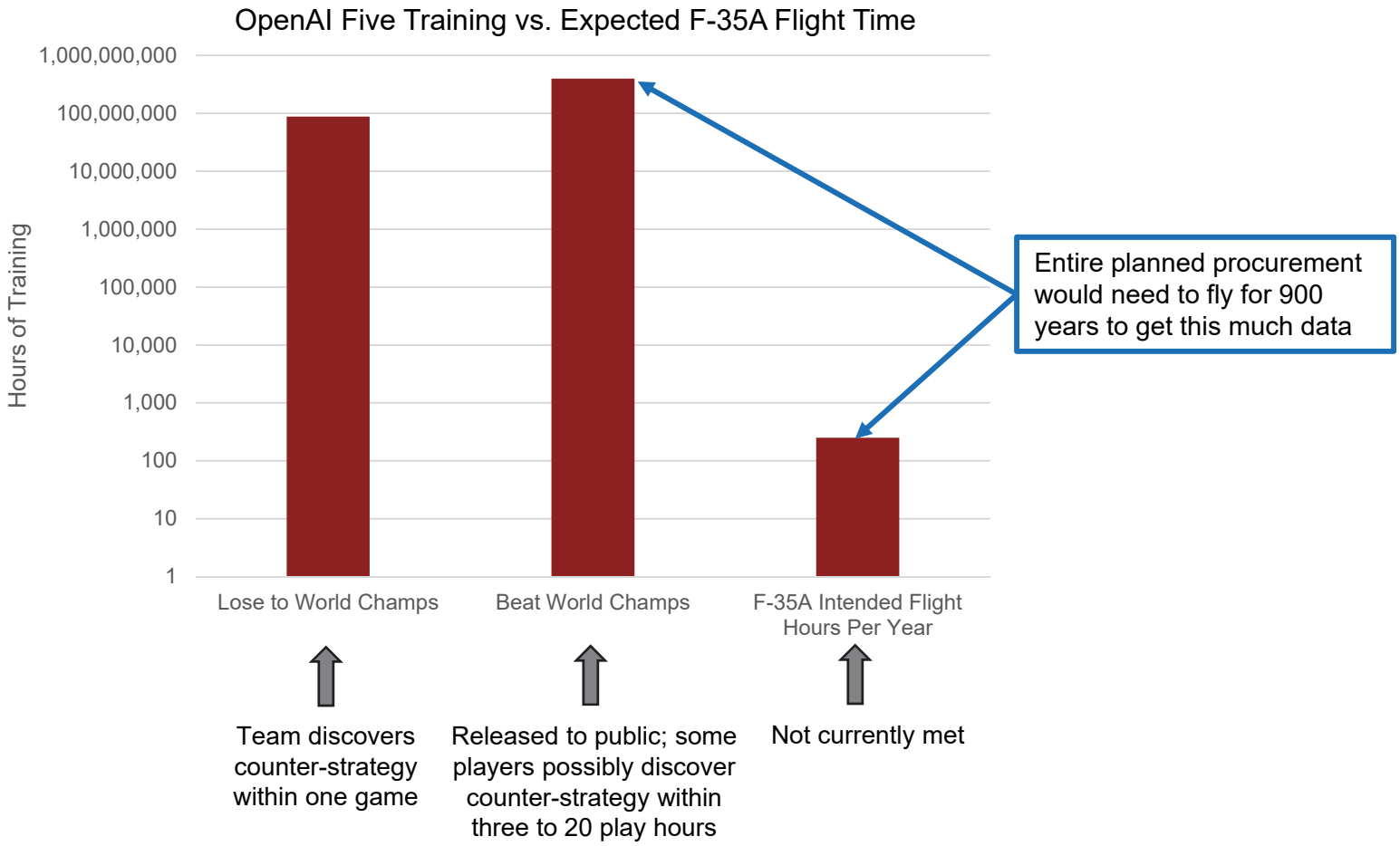
Fundamentally unsafe behaviors can be tested through graded autonomy with limited capability fielding



Operational Test (OT); Sub-capability (SC)

Some people want systems to evolve in real time.

Individual units are unlikely to meaningfully learn in real time given the state of sub-symbolic learning



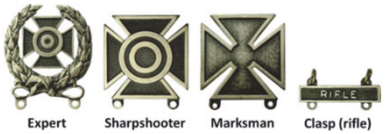
Even with batched learning,
change comes faster than T&E can handle now.

We would need to continuously recertify these systems.

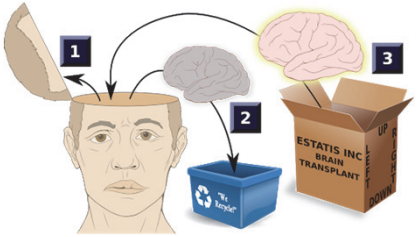
Mitigate model change challenge in online learning through O-I-D levels paradigm for recertification

O-Level: Minimal diagnostics to ensure critical decisions operating within acceptable parameters. Executable by deployed warfighters.

I-Level: More advanced diagnostics to check specific model components and recertify their functioning. Executable by experts or FSRs.



D-Level: Formal fleet-wide upgrades integrating learning and capabilities. Executable as formal testing and certification processes.



Field Service Representative (FSR); Organizational, Intermediate, Depot (O-I-D)

Discussion

Possible lessons for unintended behavior	Example
Data bias	Amazon's Hiring Recommender
Data poisoning	Microsoft's Tay
Insufficient redundancy	Max 8 crashes
Operator supervision isn't a panacea	Patriot Missile fratricides USS Vincennes Uber fatality in Arizona
Problems with ill-defined goal states need carefully chosen training and testing outcomes	Game-playing AIs (e.g., Tetris pausing)
Unexplored space with MDP-like processes can result in unintended behavior	AlphaGo Game 4 loss vs. Sedol
Not all unexpected behavior is bad	AlphaGo Game 2 Move 37

Topics for Discussion

- The need for and challenges to inference in AI
- What is the role of testing in LME?
- How do you set verifiable LME requirements?
- How do you pick test factors and outcomes for LME?
- Who should be setting requirements, factors, and outcomes?
- How do you test LME for maturing systems? For evolving ones?

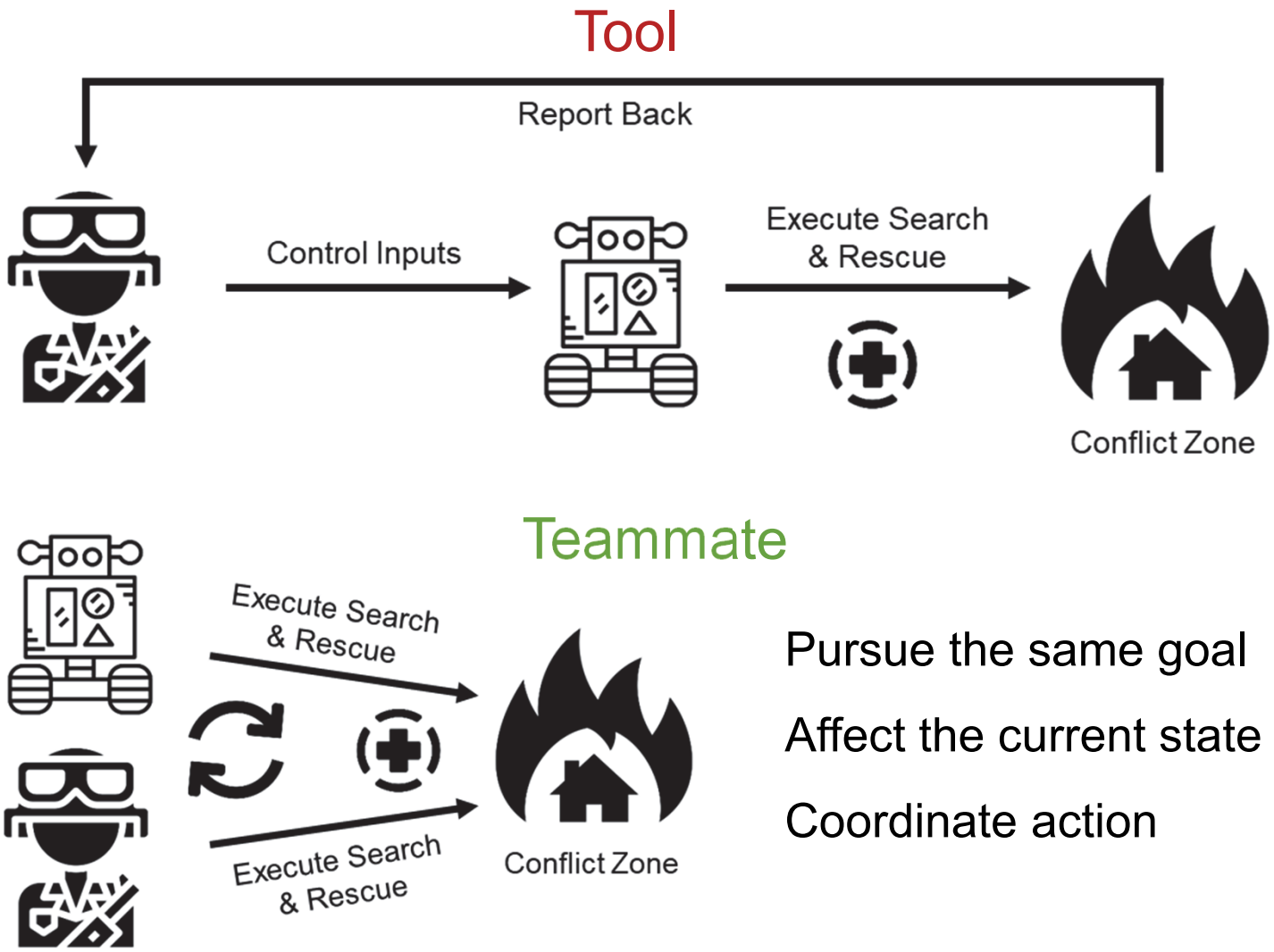
Shower/Jogging Thoughts for the Next Meeting

Emergent Behaviors

Human-machine teaming will be a major source of emergent behavior in both directions

Source: https://www.arl.army.mil/www/default.cfm?article=3244

Simply interacting with a machine doesn't make it a teammate



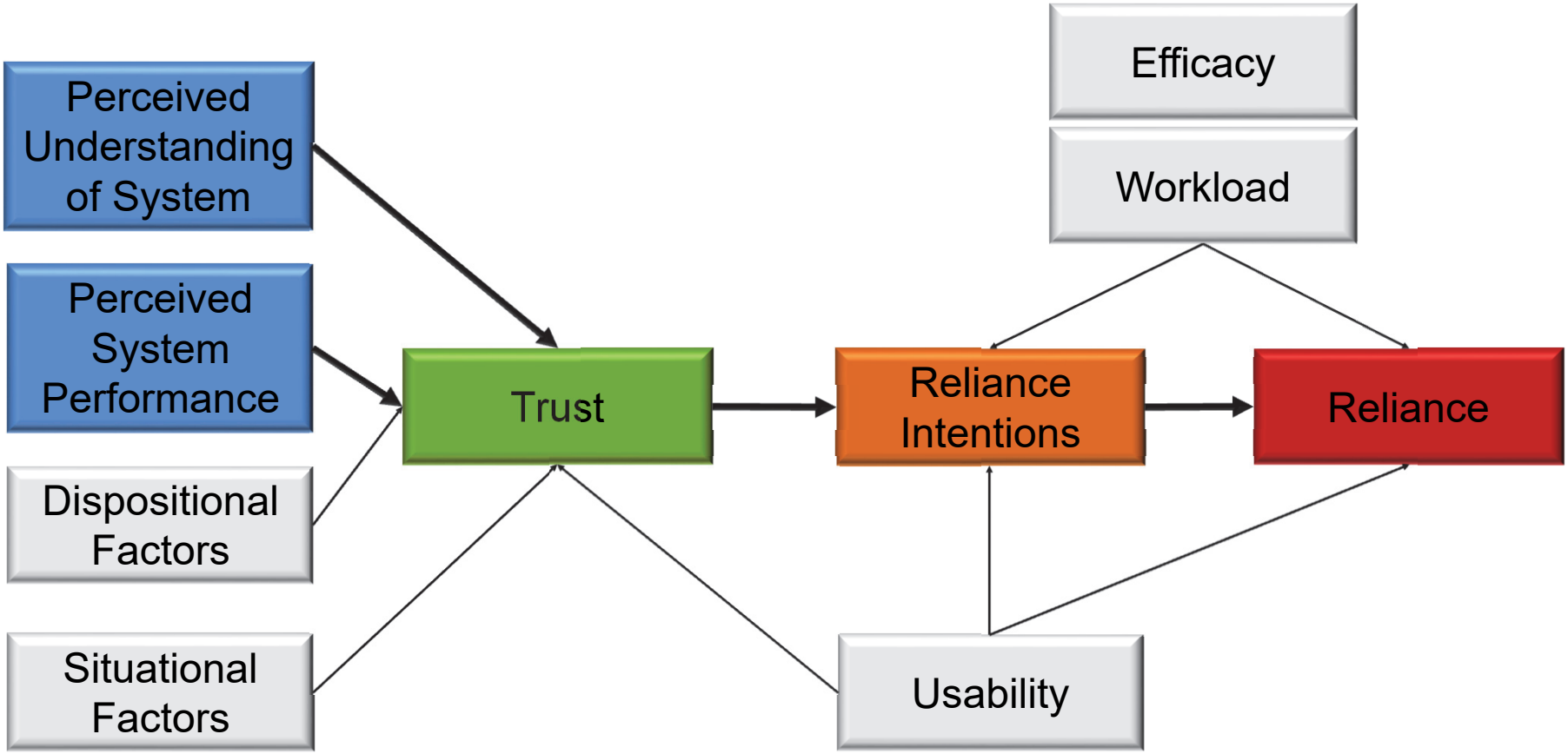
Pursue the same goal

Affect the current state

Coordinate action

Trust is a key determinant of whether operators will rely on autonomous teammates

Trust: The belief that someone or something will help you achieve your goals in a vulnerable or uncertain situation.



Backups

The DoD Ethical Principles for AI

1. **Responsible.** DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
2. **Equitable.** The Department will take deliberate steps to minimize unintended bias in AI capabilities.
3. **Traceable.** The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.
4. **Reliable.** The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
5. **Governable.** The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Possible lessons for unintended behavior	Example
Data bias	Amazon's Hiring Recommender
Data poisoning	Microsoft's Tay
Insufficient redundancy	Max 8 crashes
Operator supervision isn't a panacea	Patriot Missile fratricides USS Vincennes Uber fatality in Arizona
Problems with ill-defined goal states need carefully chosen training and testing outcomes	Game-playing AIs (e.g., Tetris pausing)
Unexplored space with MDP-like processes can result in unintended behavior	AlphaGo Game 4 loss vs. Sedol
Not all unexpected behavior is bad	AlphaGo Game 2 Move 37

Decision types in the Problem Space Hypothesis will alter what and how we test

- Executive Autonomy: Make decisions about goal states, sub-goals, problem space representations, and path constraints
  - e.g., decision-aides
- Perceptual Autonomy: Make decisions about how current problem state is defined
  - e.g., image classifiers
- Procedural Autonomy: Make decisions about next operator/procedure selected
  - e.g., Markov Decision Process

The fundamentals of performance testing are the same

All tests start with three questions:

1. What tasks are under consideration?
2. Which outcomes matter for those tasks?
3. What factors influence those outcomes?

Manned Multi-role Air Platform

Answer these questions for the system

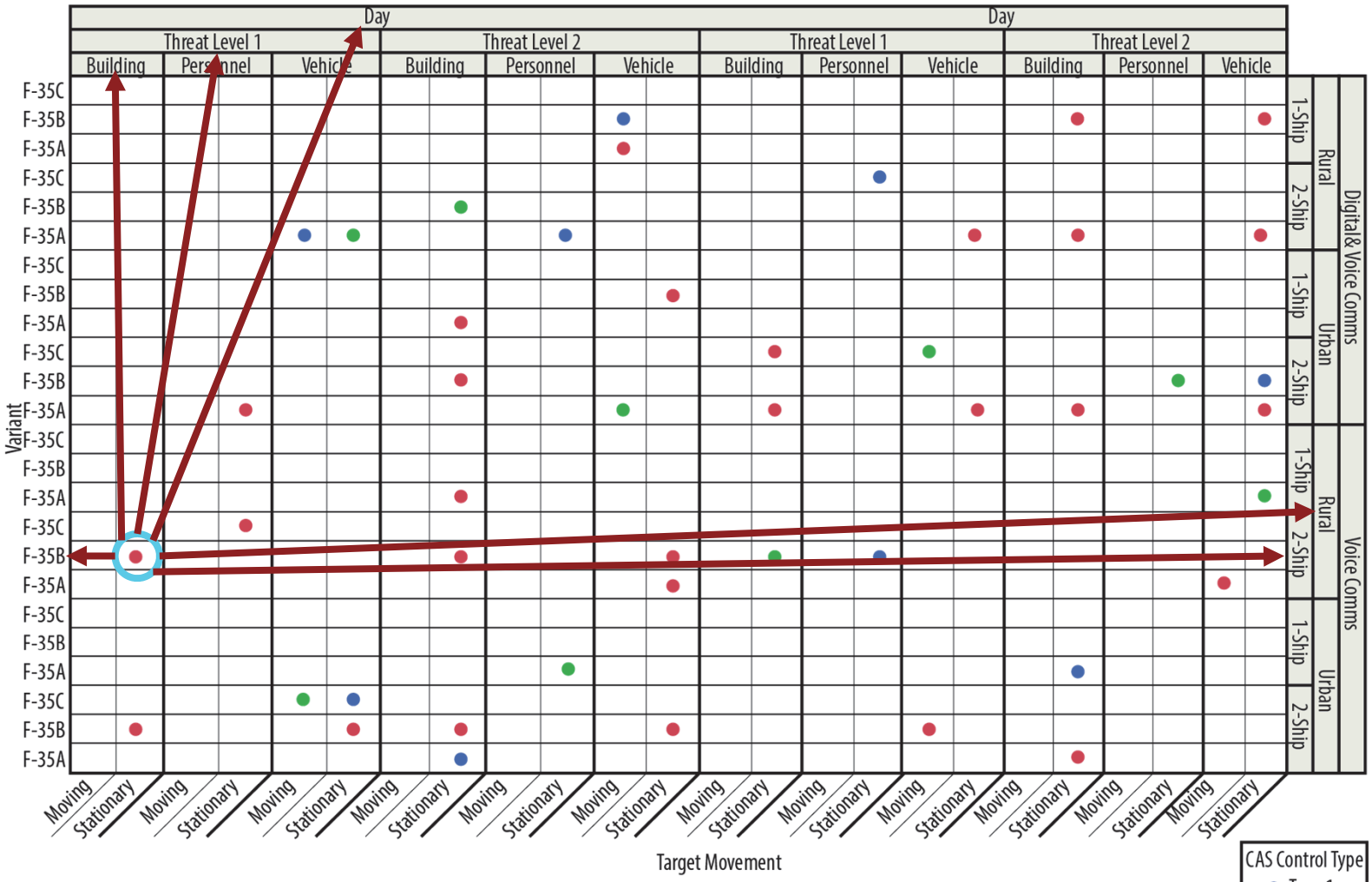
1. What tasks are under consideration?
   - E.g., Close Air Support (CAS)

2. Which outcomes matter for that task?
   - E.g., Probability of Kill; Loiter Time; Threat Exposure

3. What factors influence those outcomes?
   - E.g., System variant; CAS control level; light level; threat level; flight size; environment; target type; target speed

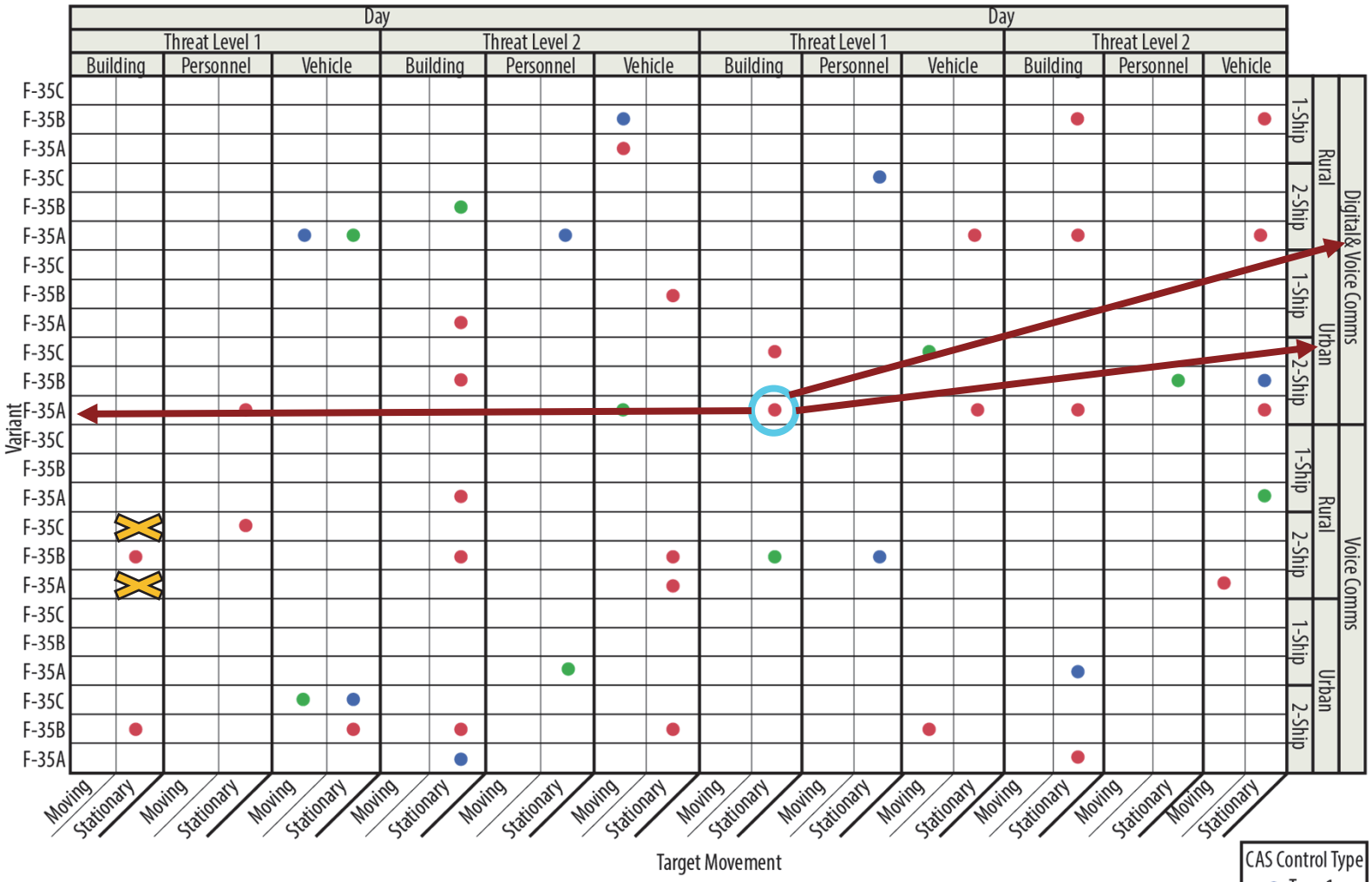
56

Distribute test points so that you can make inferences about your outcomes between your factors



Notional Close Air Support Test Points

Distribute test points so that you can make inferences about your outcomes between your factors



Notional Close Air Support Test Points

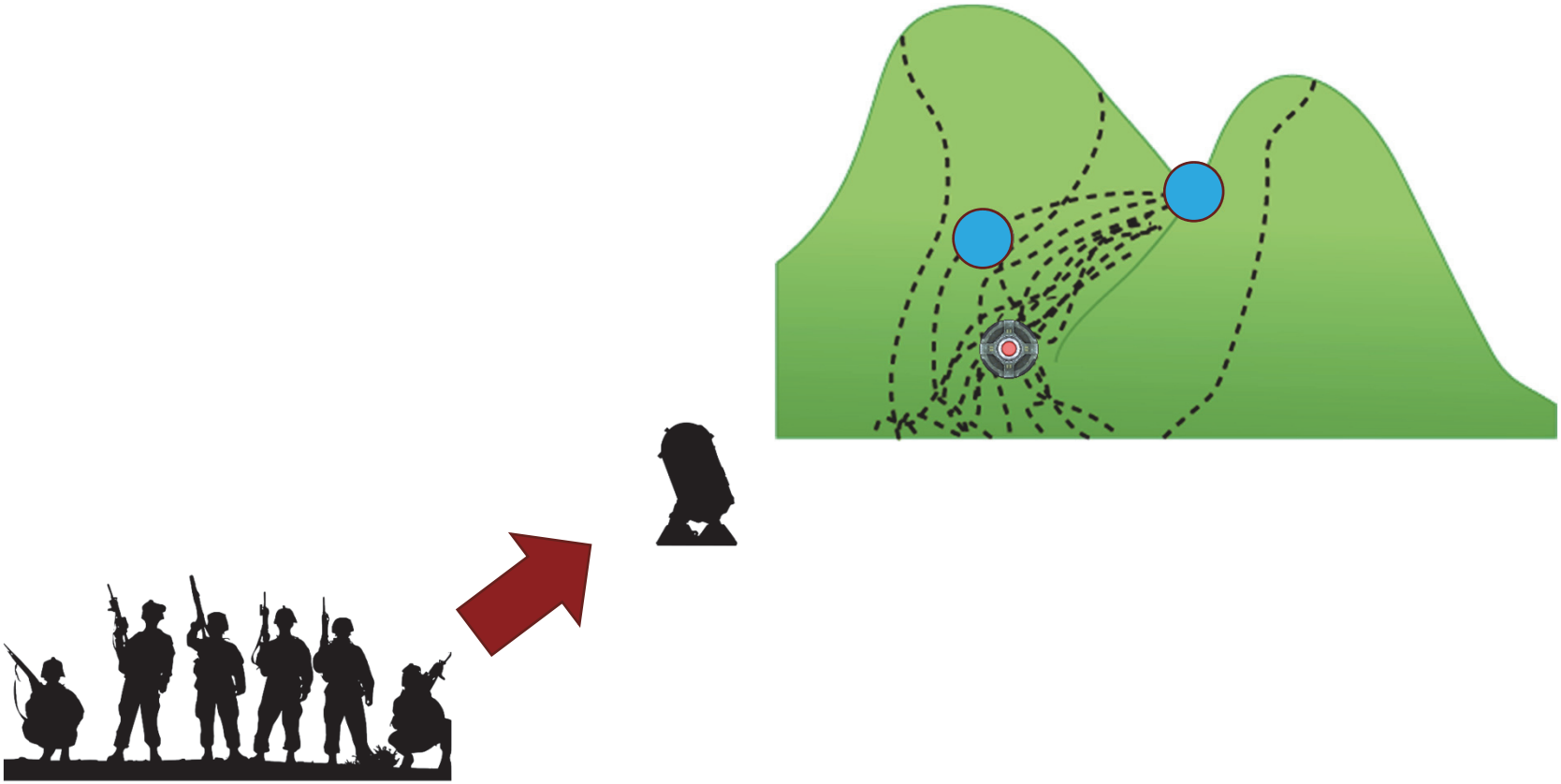
The fundamentals of performance testing are the same

- All tests start with three questions:

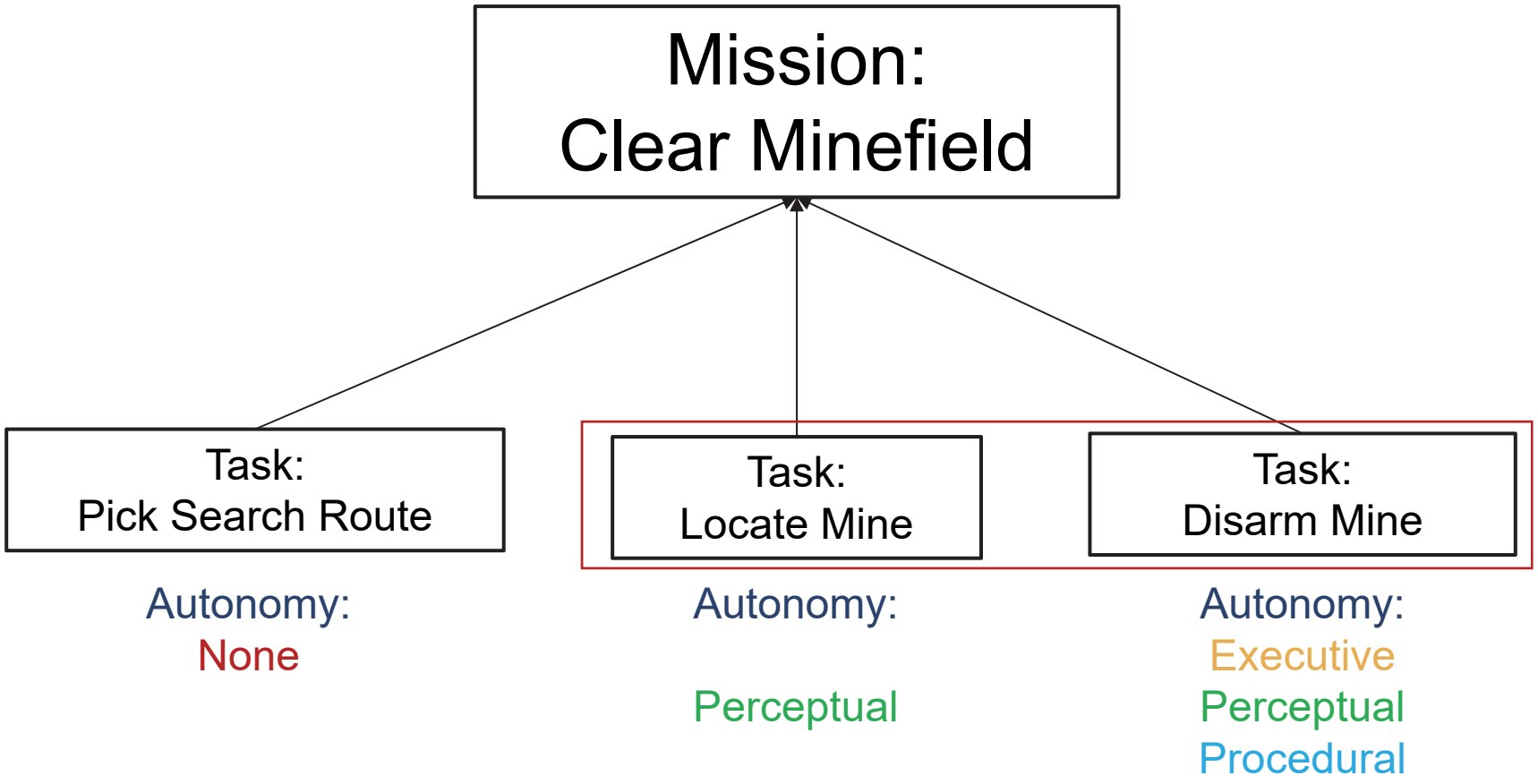
1. What tasks are under consideration?
2. Which outcomes matter for those tasks?
3. What factors influence those outcomes?

How do we answer these for decision-making systems?

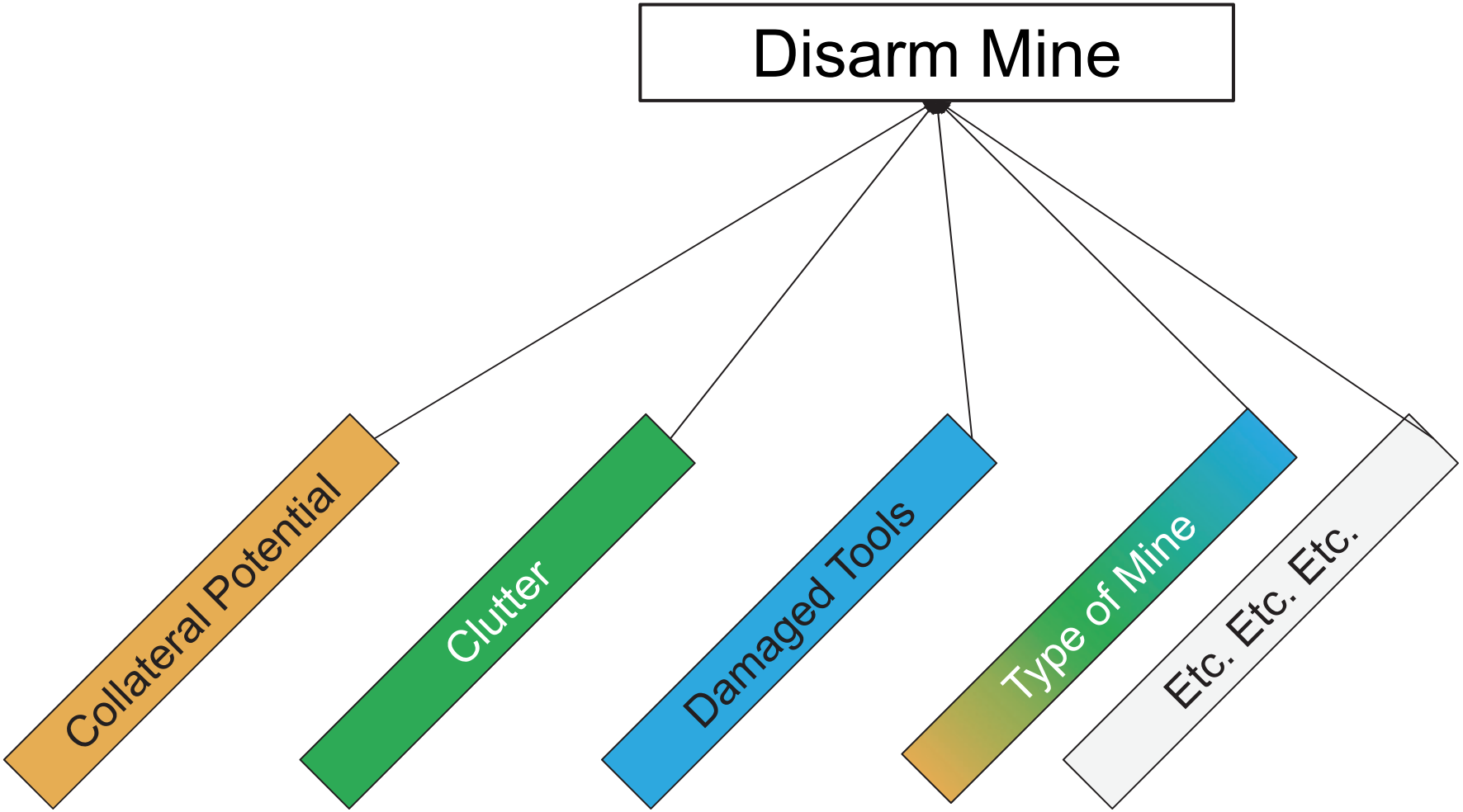
Notional System:
Robotic Reconnaissance Detonator Detector



Use task decomposition to examine the mission



Break down the task into the information dimensions that change the right answer or difficulty of getting it



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
06-2020	IDA Publication	

4. TITLE ANDSUBTITLE

Briefing to the URSA Legal, Moral, & Ethical WG:
Assuring Ethical Behavior with AI-enhanced Capabilities

5a. CONTRACT NUMBER
Separate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Daniel J. Porter (OED);

5d. PROJECT NUMBER
C9082

5e. TASK NUMBER
C9082

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
NS-D-14247
H 2020-000232

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES
Brian L. William, Project Leader

14. ABSTRACT
The Defense Advanced Research Projects Agency (DARPA) convened a working group to examine the Legal, Moral, & Ethical (LME) considerations for artificial intelligence (AI) enhanced capabilities (AIECs). Part of this consideration is how you would provide assurance that a system is LME compliant. OED has developed a framework for providing assurance more generally for AI-enabled systems, and the LME working group invited the authors to brief them on this topic.

15. SUBJECT TERMS
Artificial Intelligence (AI); TARTA; test, evaluation, verification, and validation (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**			Daniel J. Porter
Unclassified	Unclassified	Unclassified	Unlimited	71	**19b. TELEPHONE NUMBER** *(include area code)* 703-578-2869

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI std. Z39.18