INSTITUTE FOR DEFENSE ANALYSES

Briefing to the Air Force Scientific Advisory Board: T&E Contributions to Avoiding Unintended Behaviors in Autonomous Systems

Heather M. Wojton, Project Leader

Daniel J. Porter

February 2020

Approved for public release.

Distribution is unlimited.

IDA Document NS D-12078

Log: H 2020-000049

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 contract HQ0034-19-D-0001, Task C9089 "Trust in Automation." 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 of Dean Thomas and Mark Herrera from the Operational Evaluation Division, and David Sparrow from the Science & Technology Division.

For more information:
Heather M. Wojton, Project Leader
hwojton@ida.org • 734-845-6811

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-12078

Briefing to the Air Force Scientific Advisory Board: T&E Contributions to Avoiding Unintended Behaviors in Autonomous Systems

Heather M. Wojton, Project Leader

Daniel J. Porter

Executive Summary

A. Background

The Air Force Scientific Advisory Board (AFSAB) is conducting a study on Understanding and Avoiding Unintended Behaviors in Autonomous Systems. IDA has been developing a framework for providing assurance for autonomous systems, and the AFSAB 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 Air Force Scientific Advisory Board: T&E Contributions to Avoiding Unintended Behaviors in Autonomous Systems

Dr. Daniel J. Porter

Dr. Heather M. Wojton

February 10, 2020

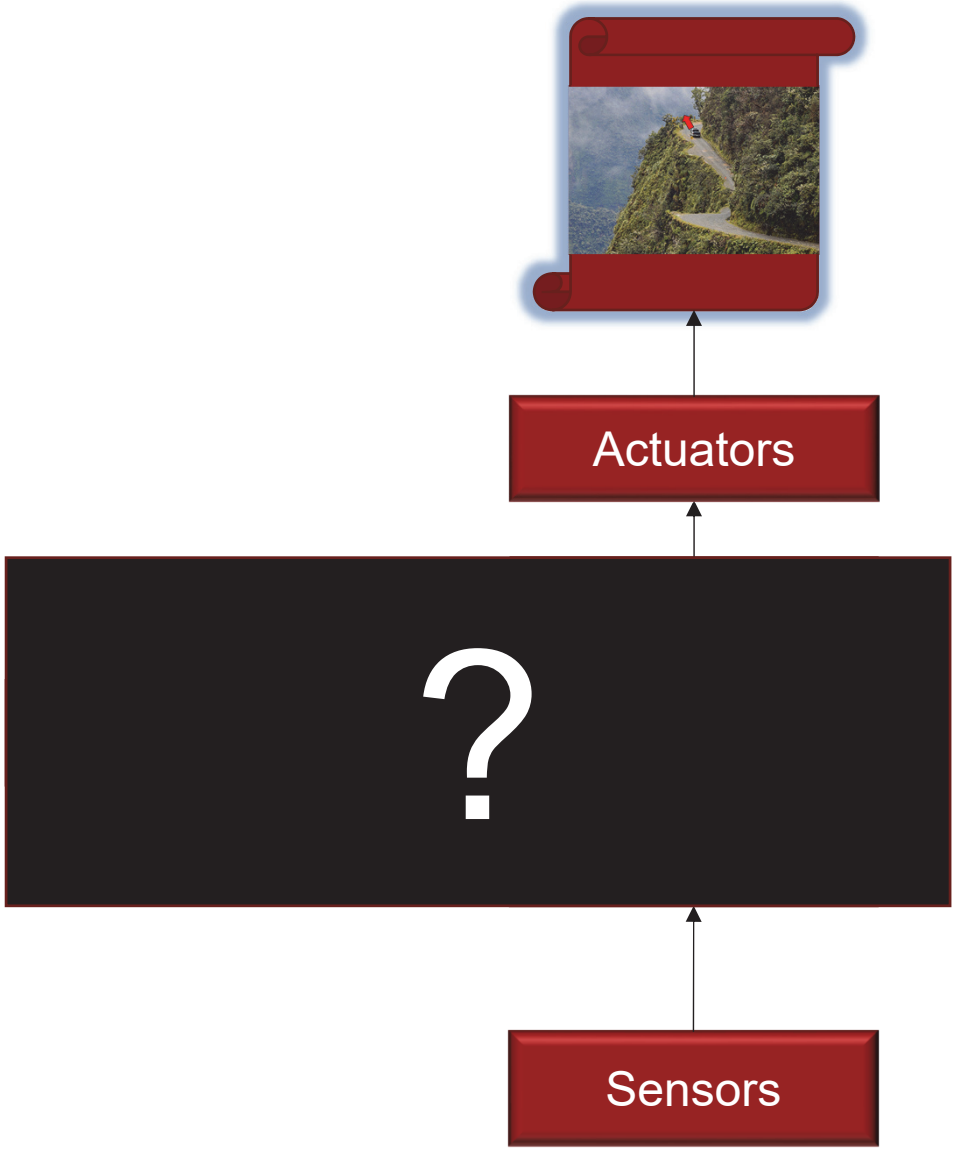
Institute for Defense Analyses

4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

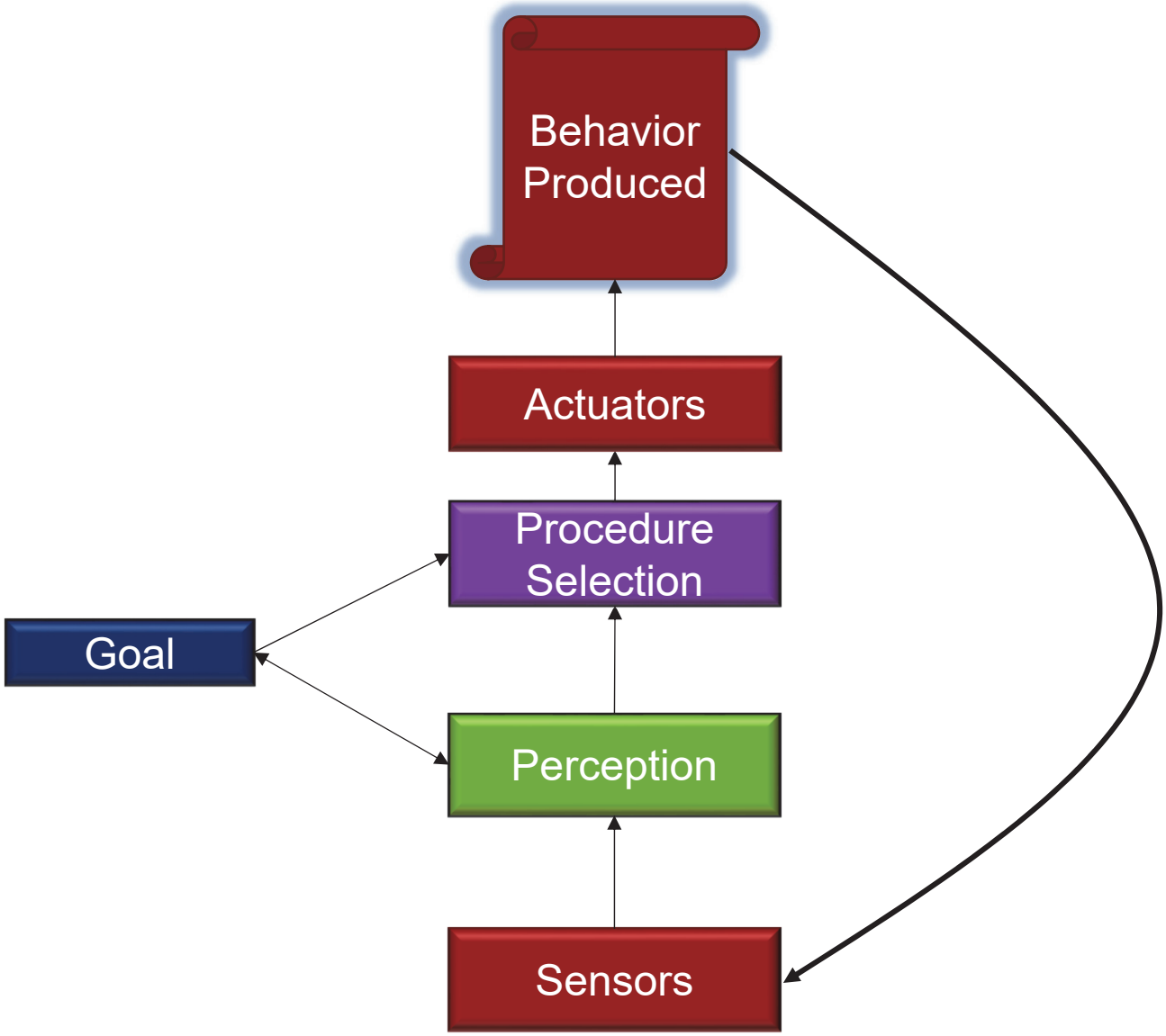
The ability to make valid inferences is the best defense against unintended behaviors.

Inferring behavior requires understanding the decisions that causally drive those behaviors

We cannot generalize behavior from black boxes



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



Diagnosing unintended behavior will require unobtrusive instrumentation on decision processes

"Cognitive Instrumentation"

We need to instrument these across entire system lifecycle

Actuators

Procedure Selection

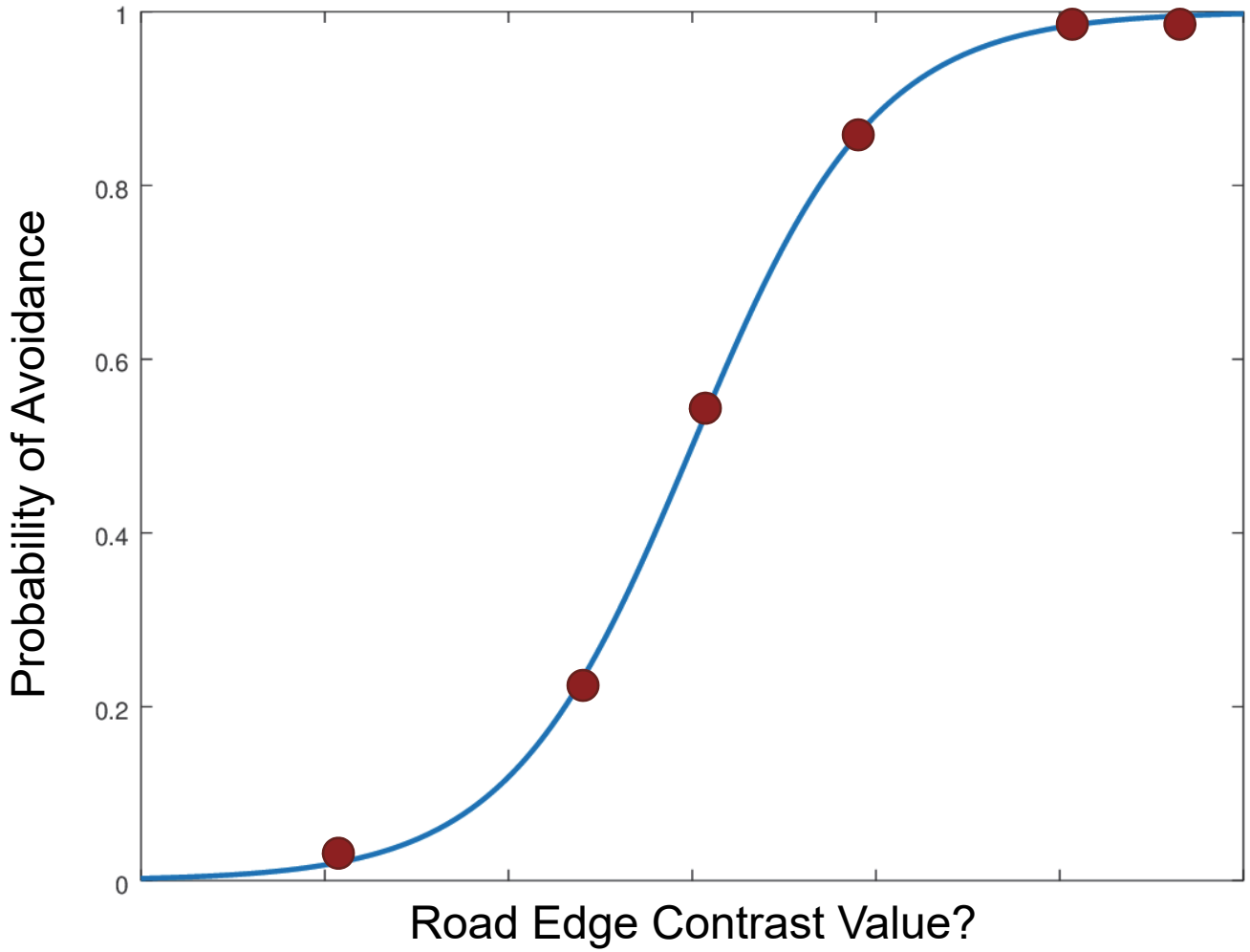
Goal

Perception

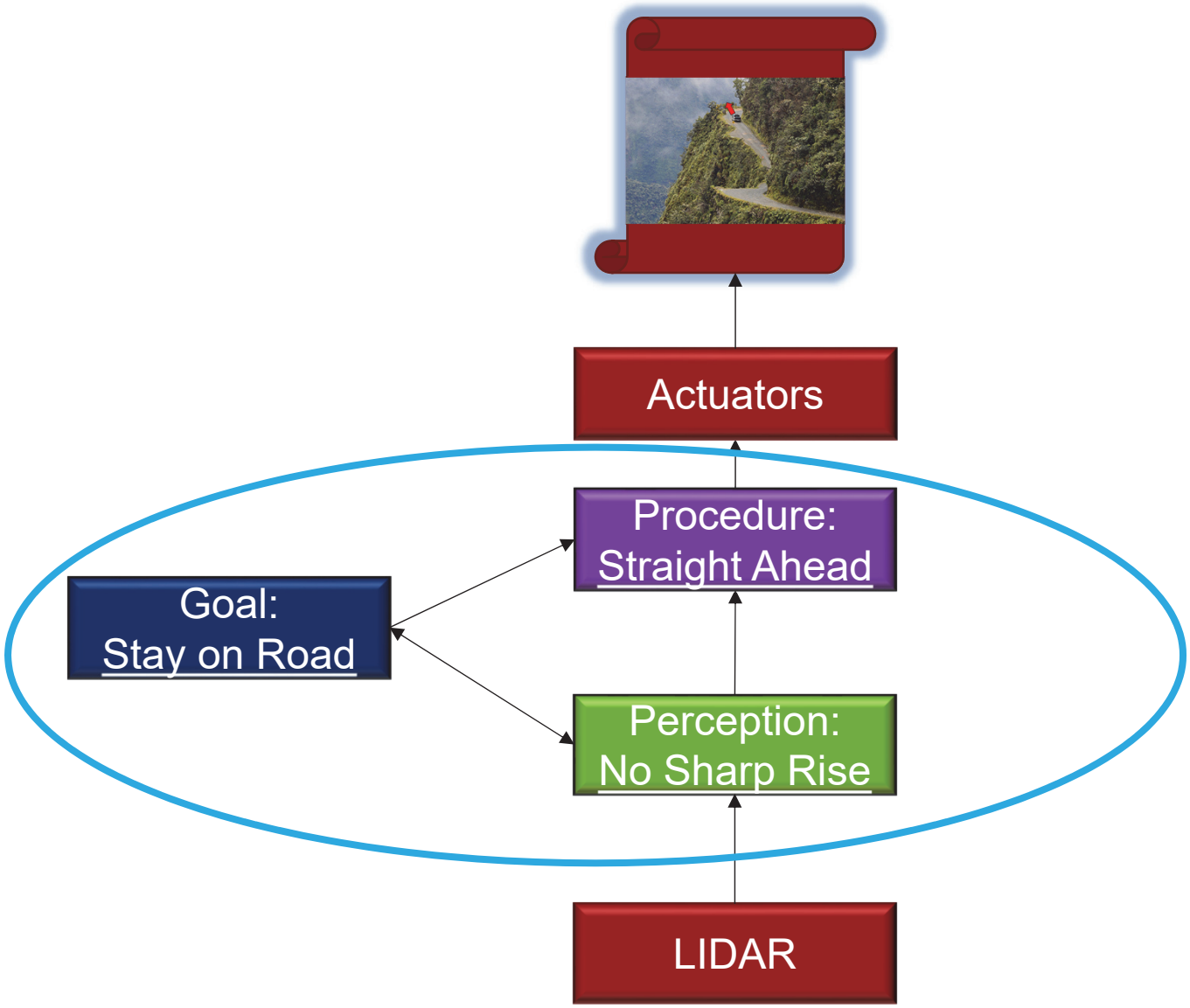
Sensors

We're used to instrumenting these in test

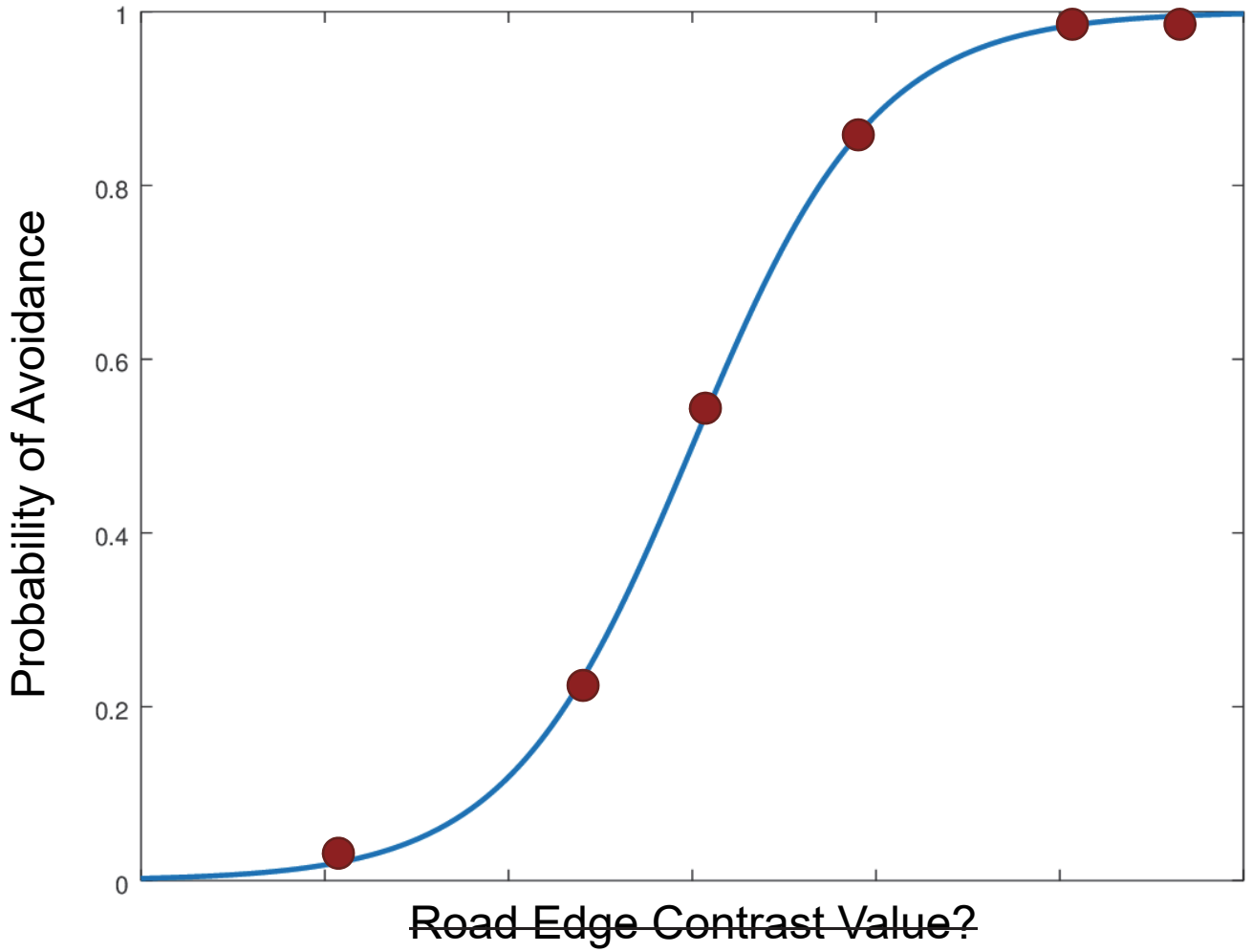
We ultimately want to validly generalize across information dimensions to avoid unintended behaviors



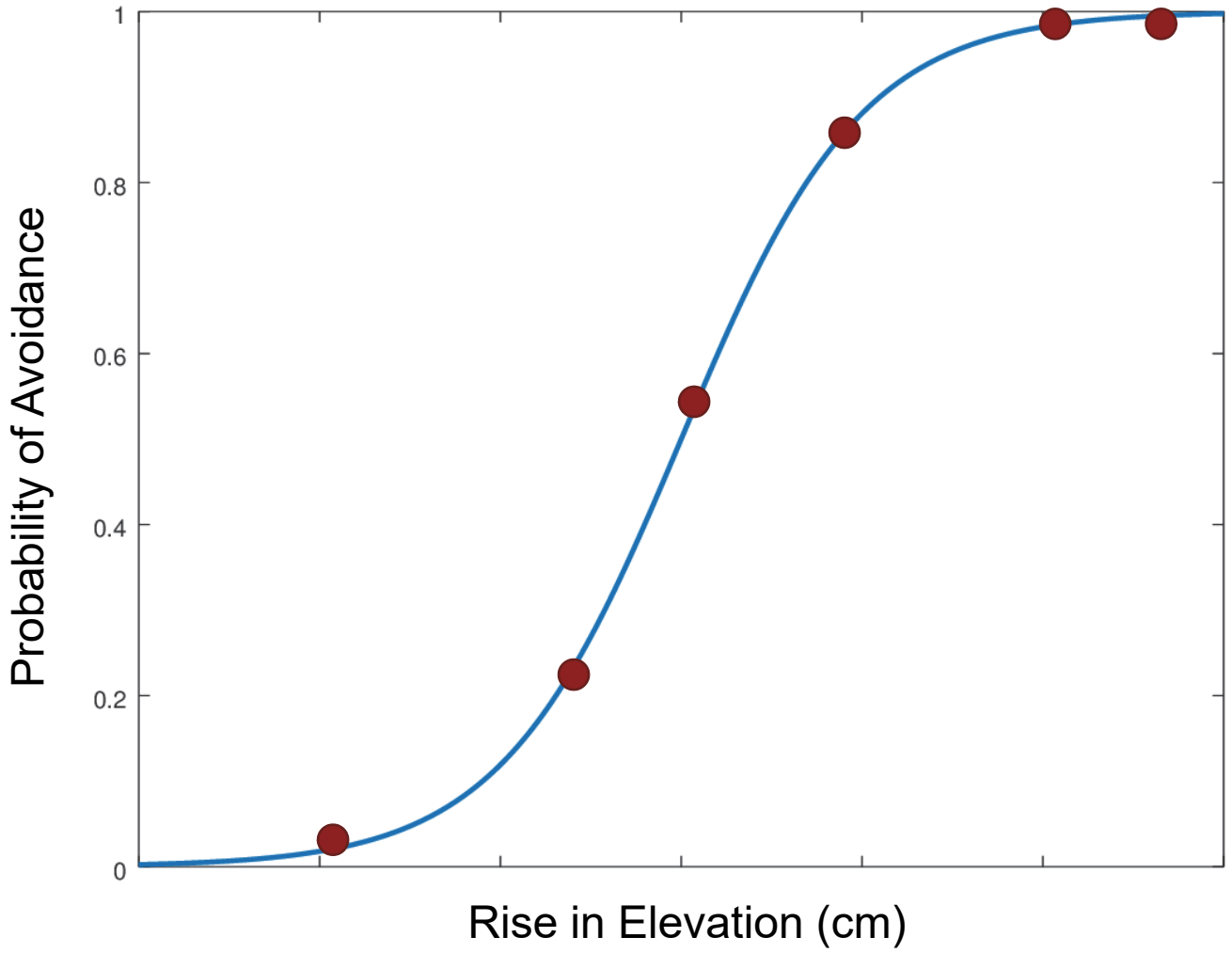
We have to obtain, verify, validate, and accredit models of system decision making



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

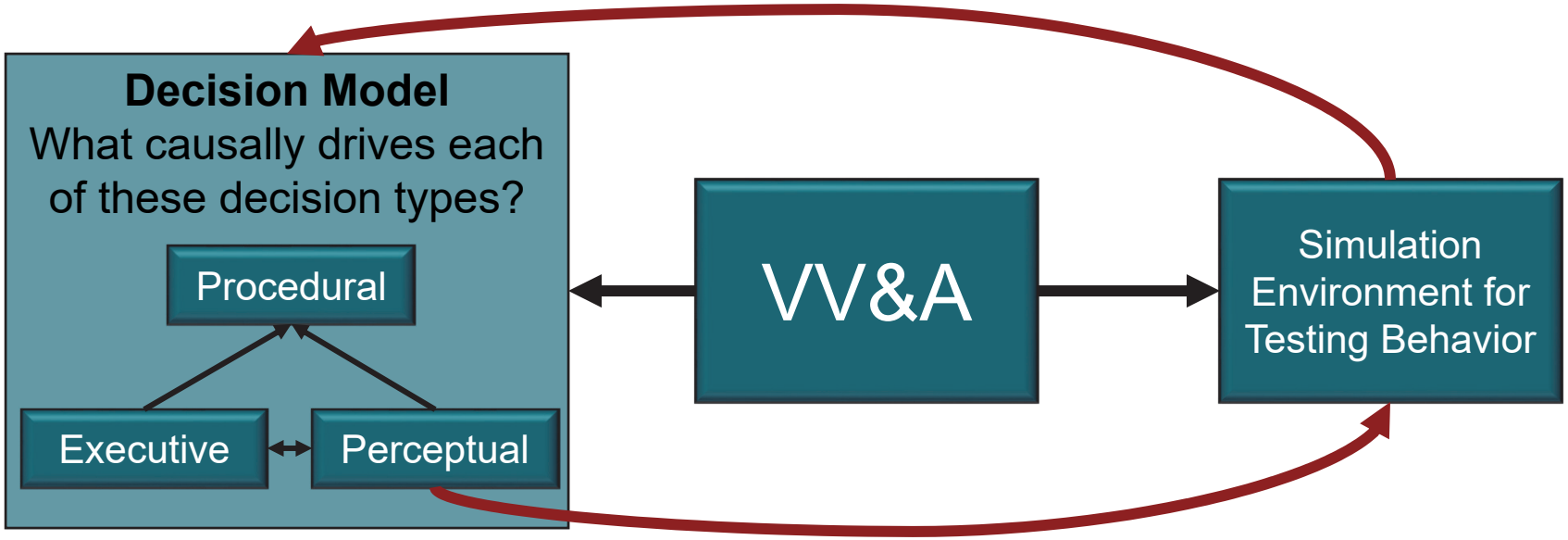


We need to ensure the information dimensions varied in test are the causal drivers and not just correlated



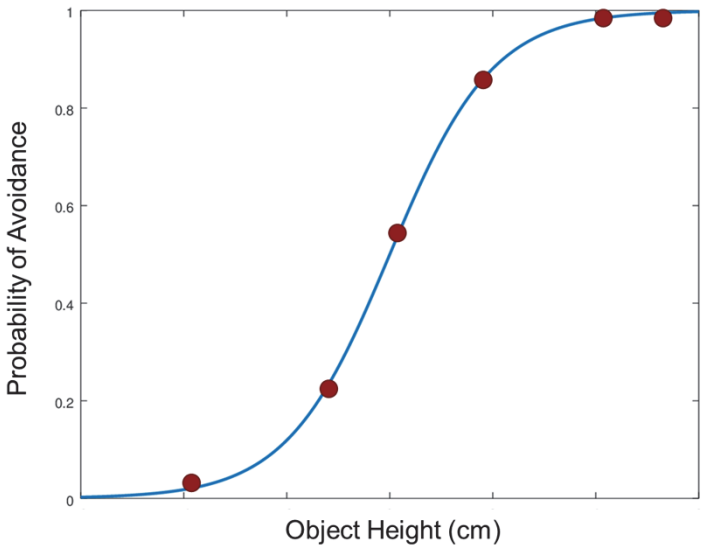
How to obtain, verify, validate, and accredit system decision models

VV&A needs to happen for more than one thing

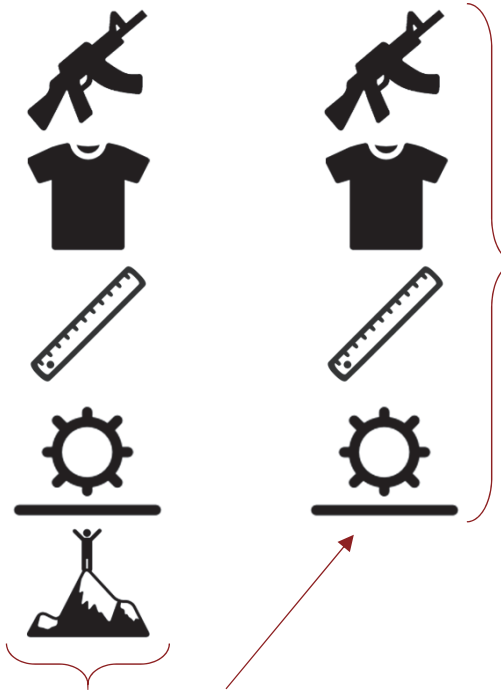


Decision model goal is inference over test factors but sub-goal is VV&A of behavioral simulation environment

Main Goal:
Infer between and beyond performance observations



Sub-goal:
VV&A Behavioral Simulation

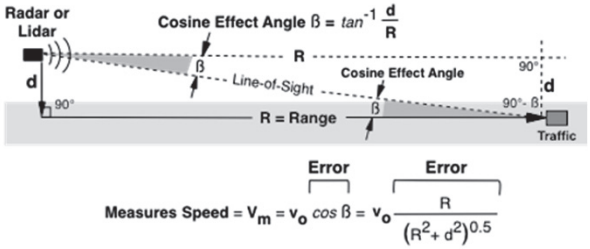


Environment VV&A
Adequate representations?

Are these in here?
Decision Model OVVA

Sensor physics can be valid without the environmental features being valid and representative

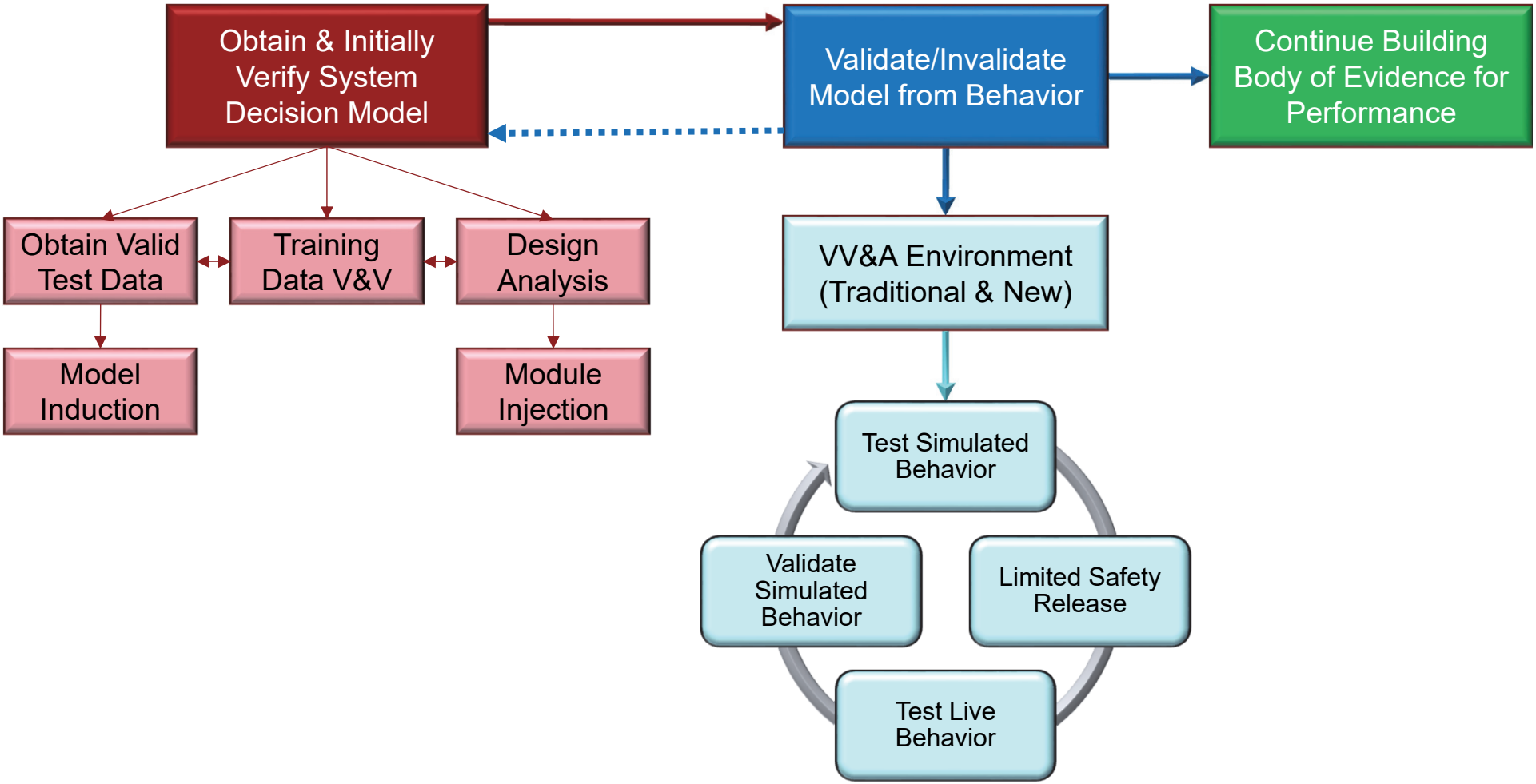
θ is a causal driver of threat perception



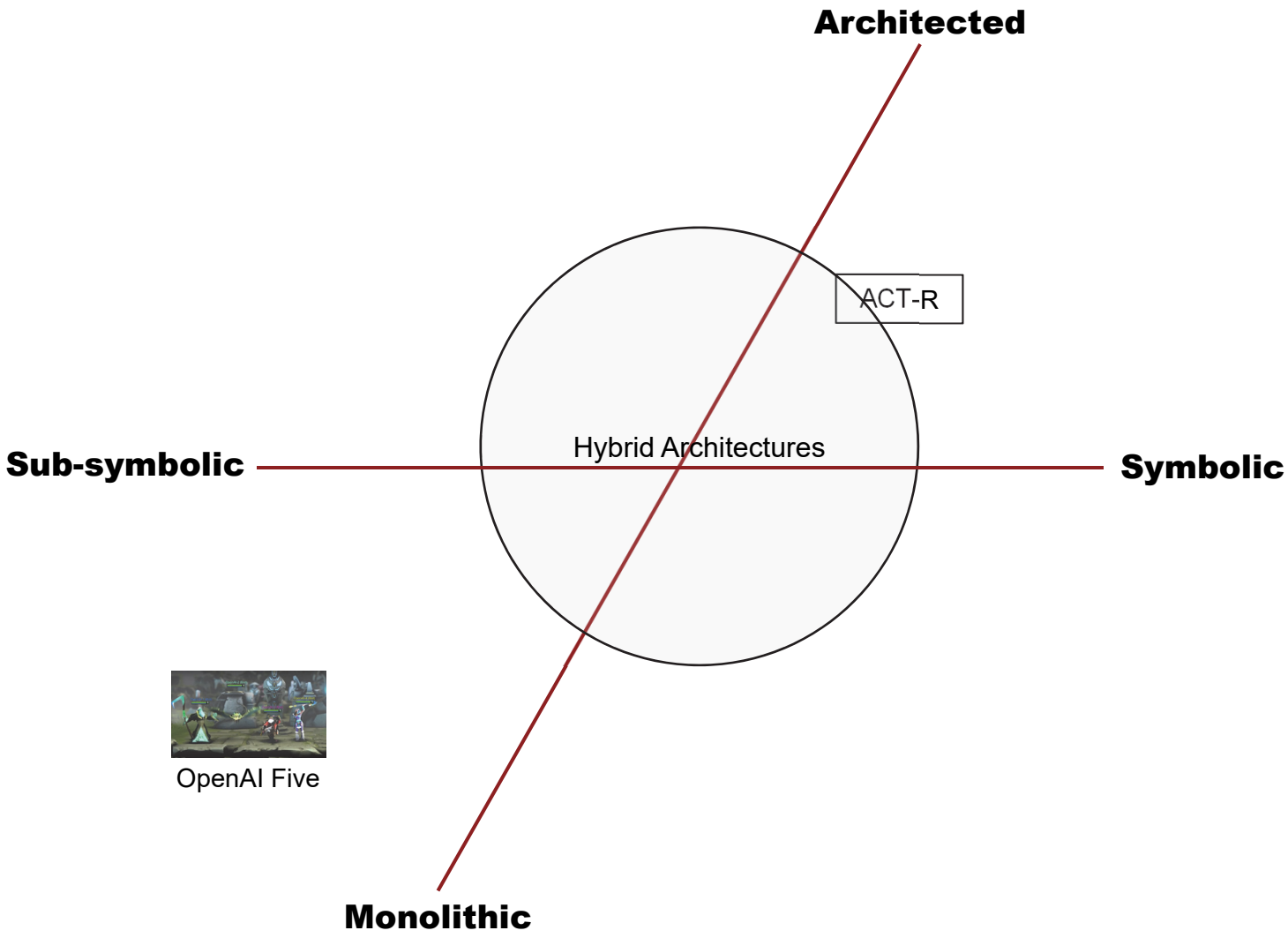
Behavioral sim doesn't vary barrel angle

OVVA requires iterative test and evaluation

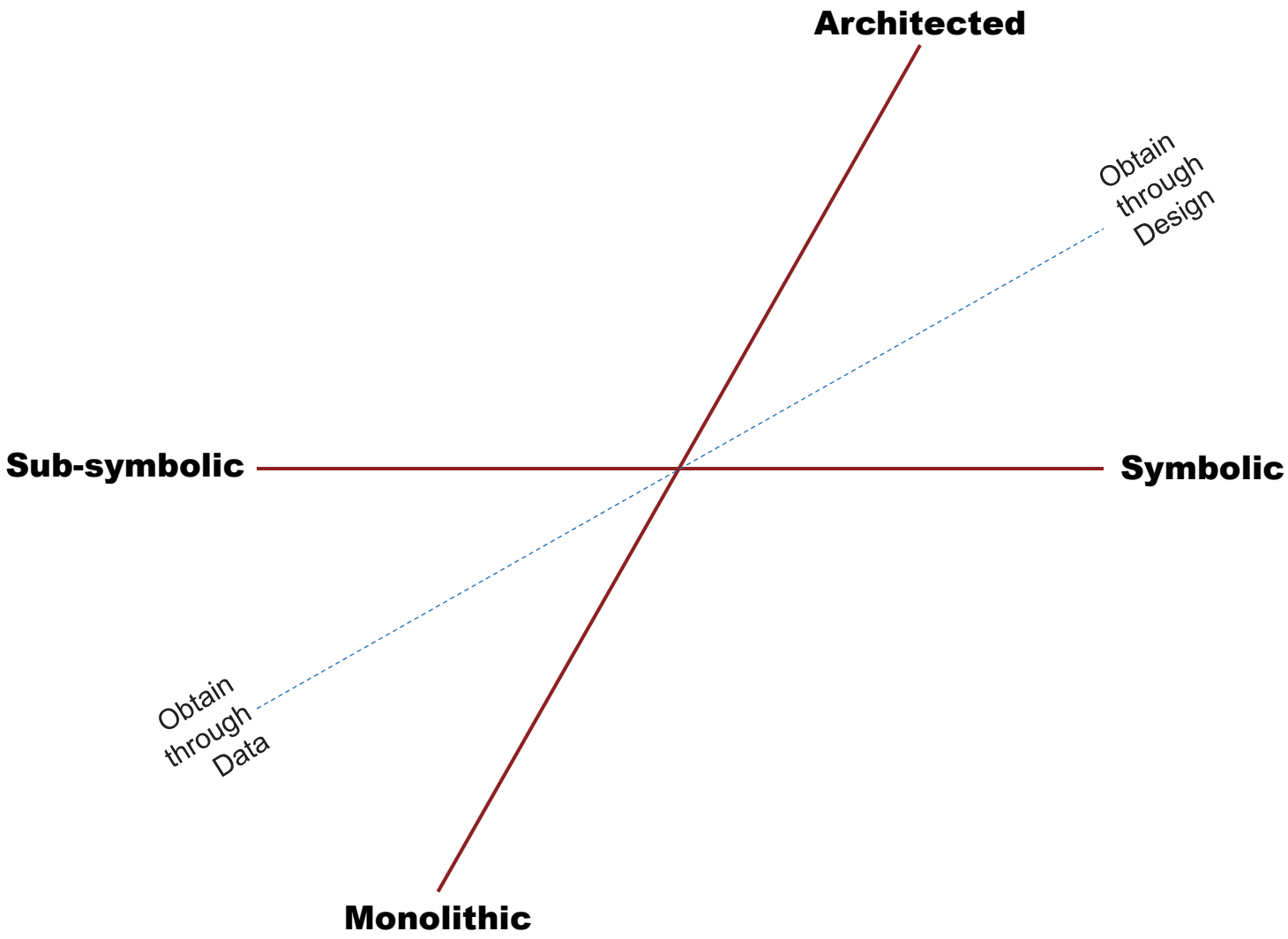
OVVA requires iterative test and evaluation



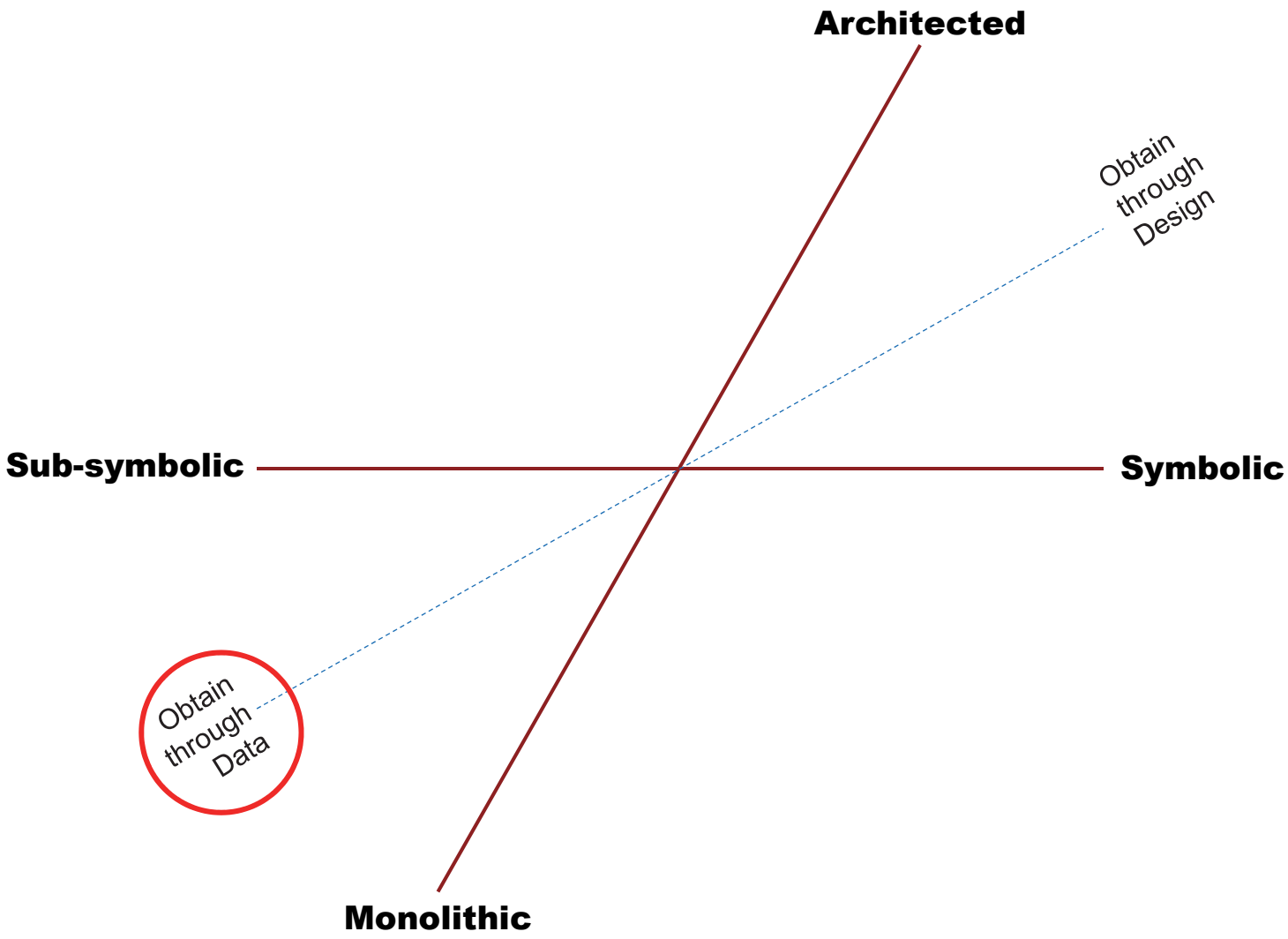
System design alters how to obtain decision model



System design alters how to obtain decision model

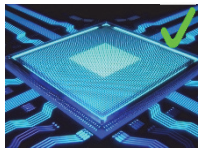


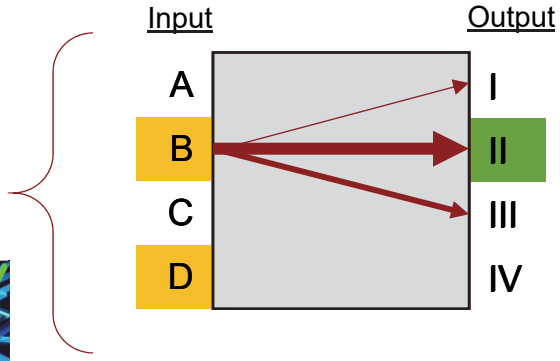
System design alters how to obtain system model



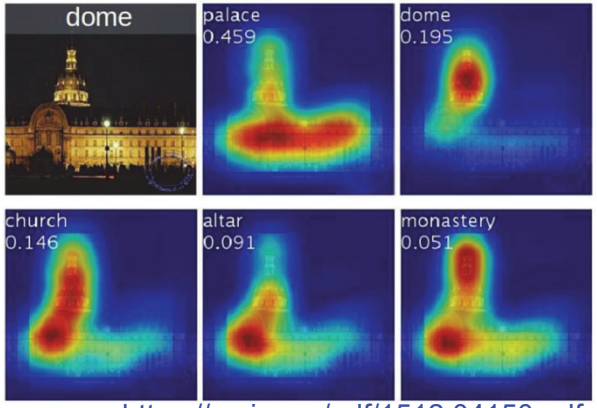
Model induction has promising data-driven techniques, but may be insufficient for embedded full autonomy

Assumption:
Have valid inputs



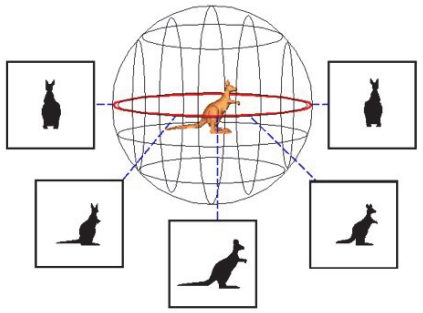


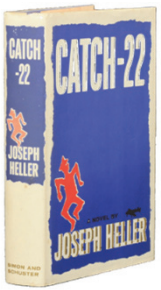
E.g., Saliency Mapping



https://arxiv.org/pdf/1512.04150.pdf

Full autonomy can change the information acquired

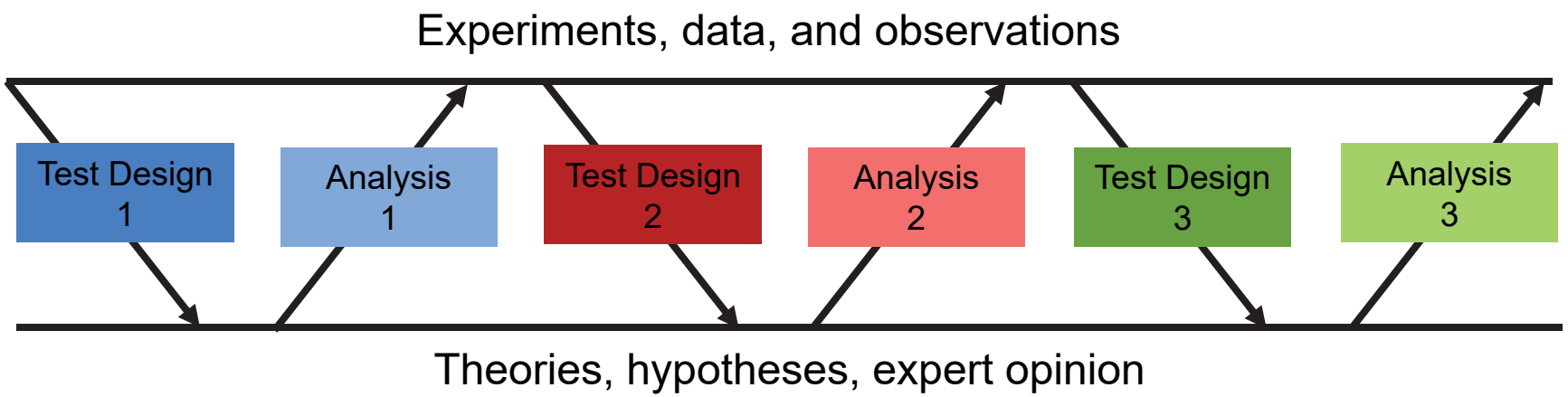


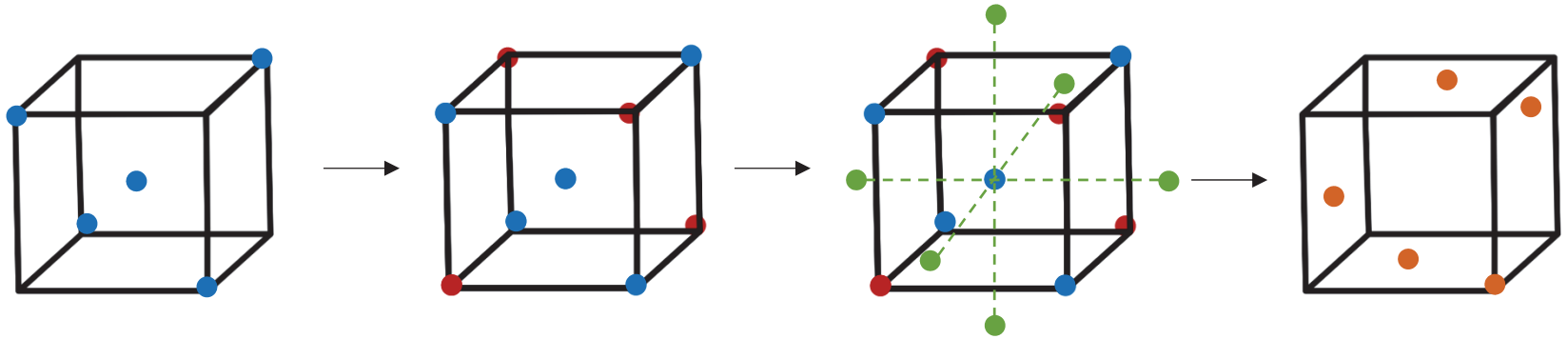


Need model to VV&A sim

Need sim for safety release

Sequential experimentation is (likely) the most efficient method for model induction

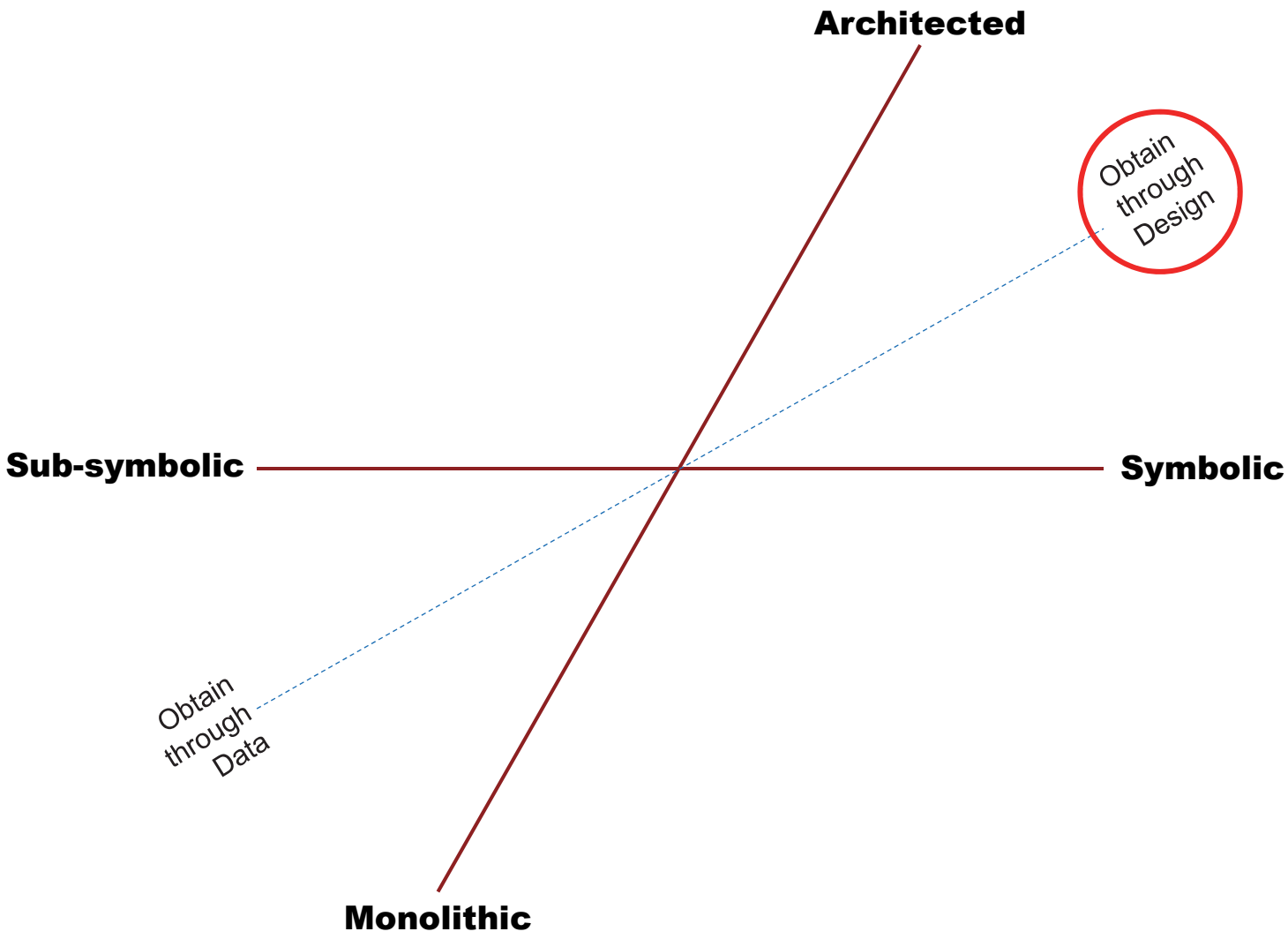




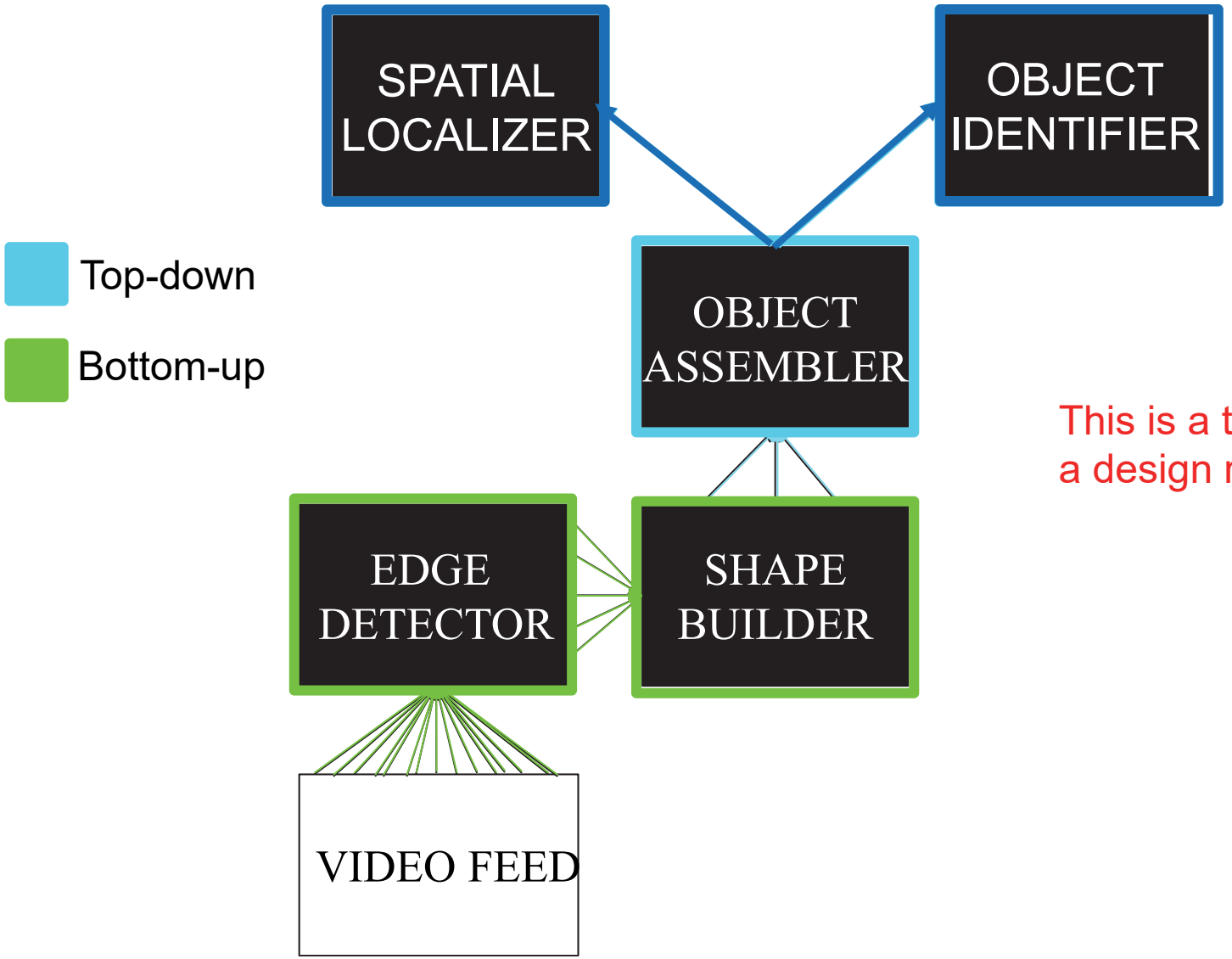
Sub-symbolic, monolithic systems will demand much greater quantities of data to obtain decision models.

These data may be expensive for both time and resources.

System design alters how to obtain system model

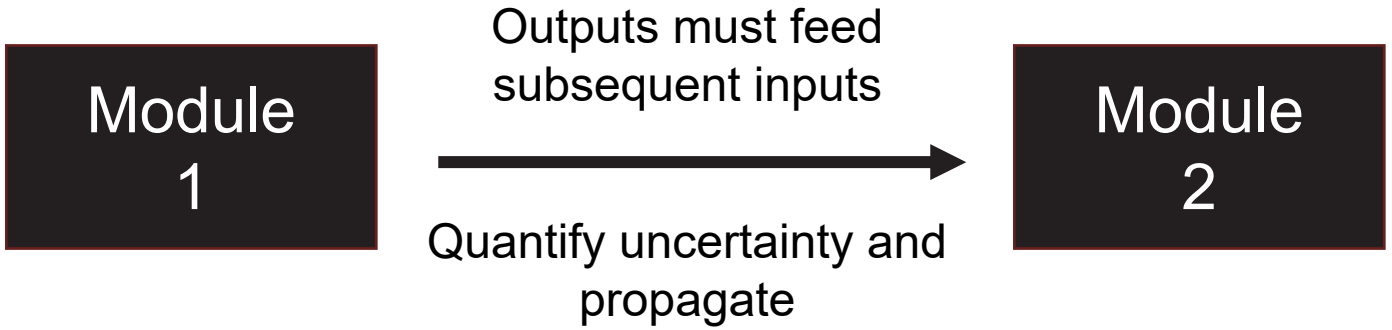


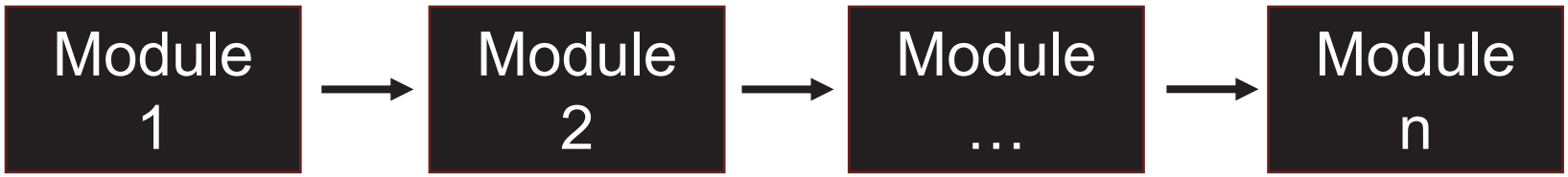
Modular architectures' decision models can be initially verified through cascading compositional verification



This is a toy example, not a design recommendation

Bayesian network models can quantify uncertainty in decision making across distributed modules

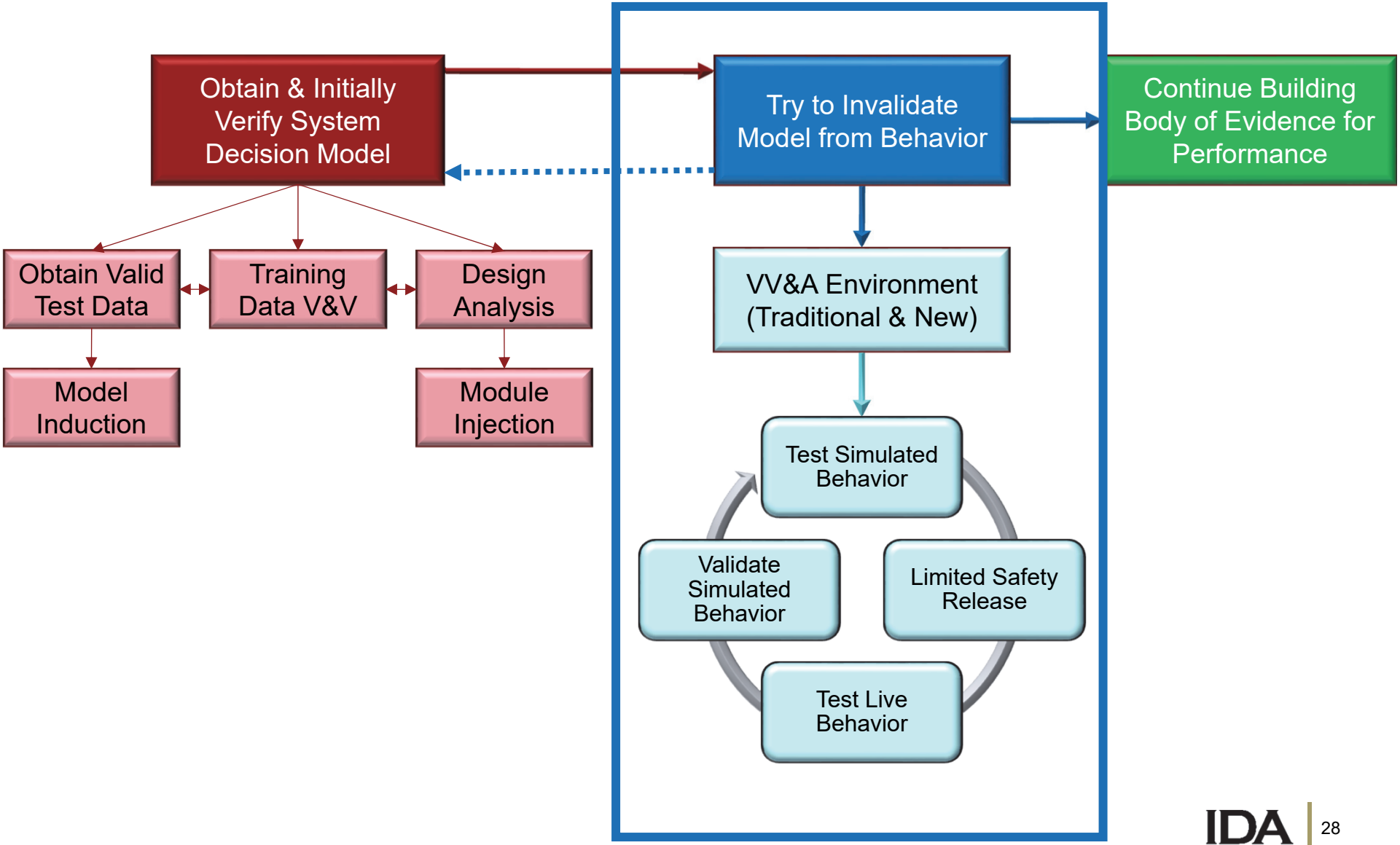




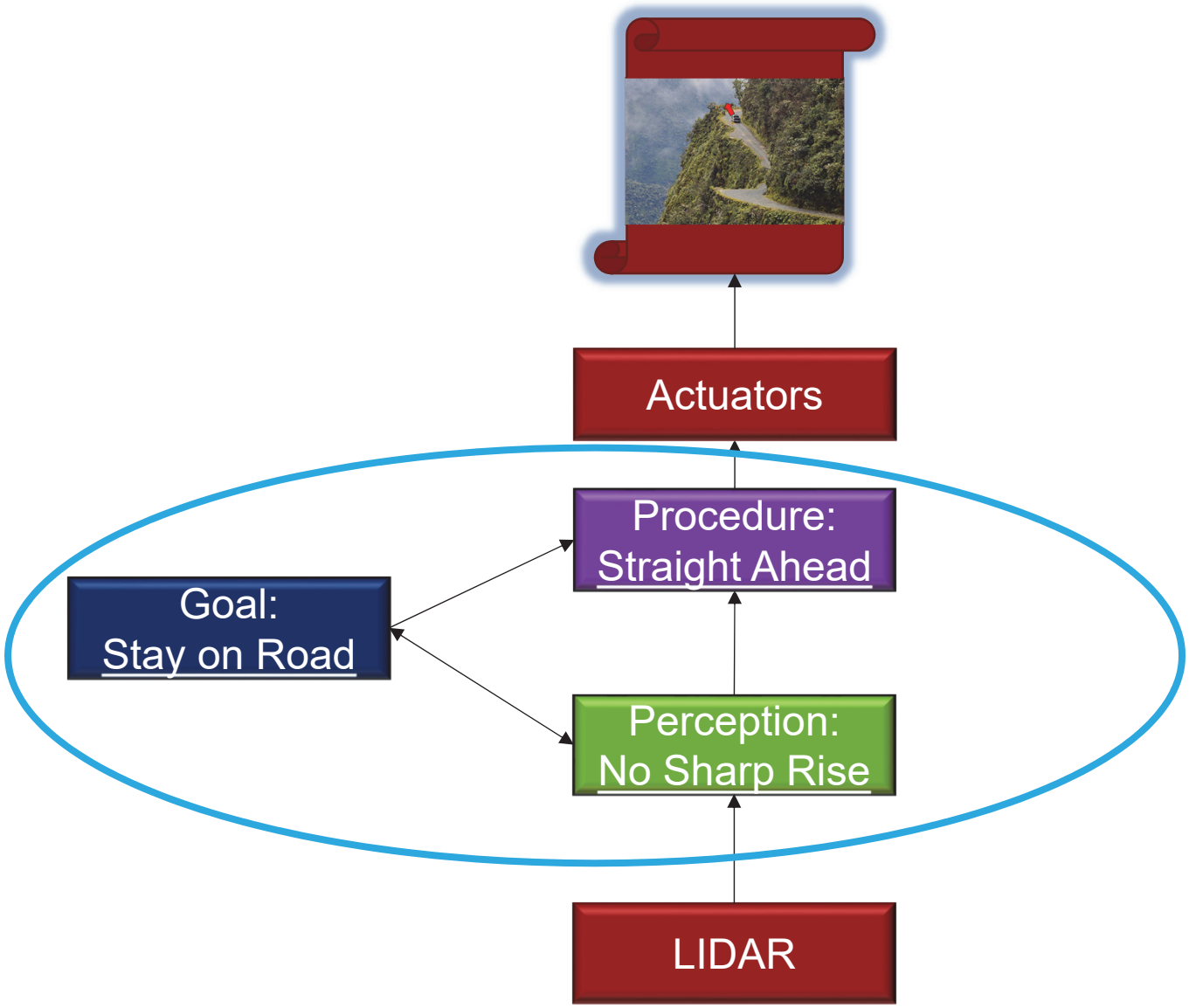
Propagating uncertainty across multiple modules provides uncertainty estimates in all or part of the decision model, supporting verification

Modularized architectures make it easier
to obtain and verify system decision models.

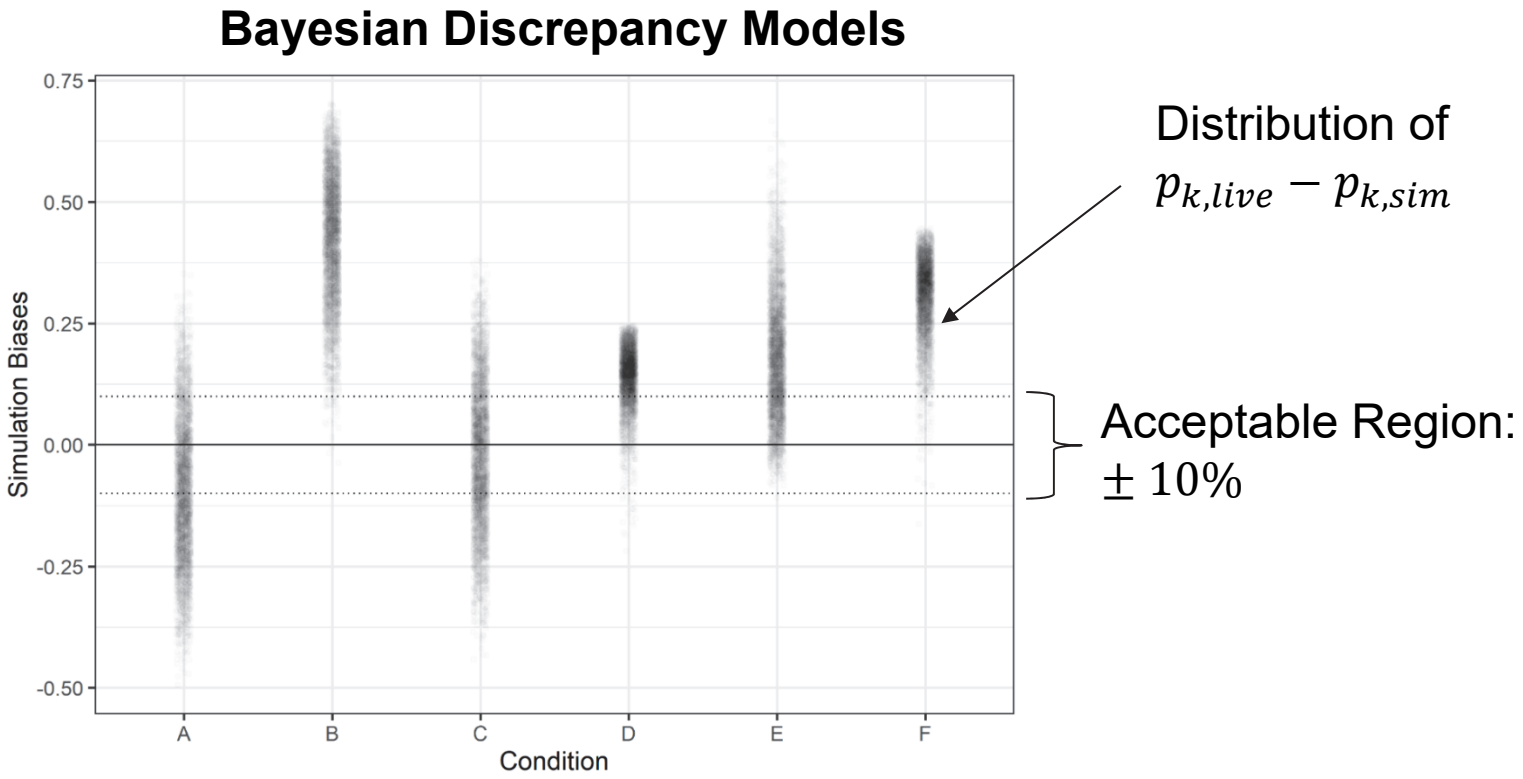
Once you have a model, spiral through risk trying to invalidate that model with simulated and live behaviors



Test scenarios that could disprove your model



A range of statistical methods exist to determine the extent to which models match reality



Use a simulation of the decision model to predict the live outcome at each condition

Model the *discrepancy* between the live and simulation outcomes

Diagnosing Behavior

Recommendation: Include instrumentation and data infrastructures as formal requirements across system lifecycles.

We will need to understand the causes of unintended behaviors that occur from early development until the end of its operational life.

Diagnosing unintended behavior will require unobtrusive instrumentation on decision processes

"Cognitive Instrumentation"

We need to instrument these across entire system lifecycle

Actuators

Procedure Selection

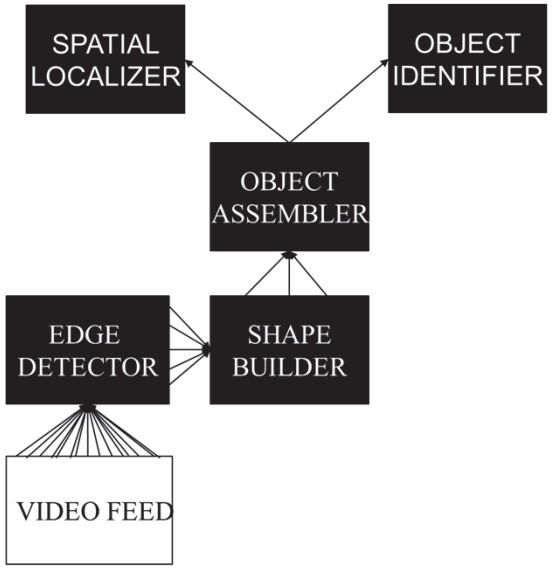
Goal

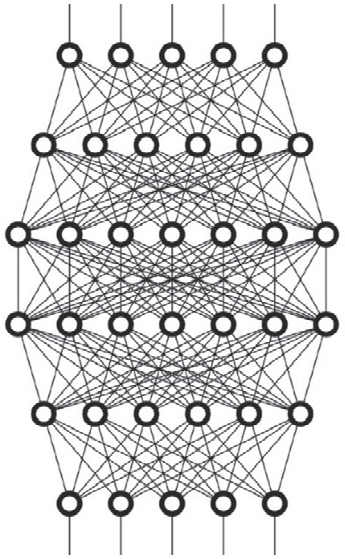
Perception

Sensors

We're used to instrumenting these in test

What, where, and how we instrument will depend on system design, current transparency, and maturity



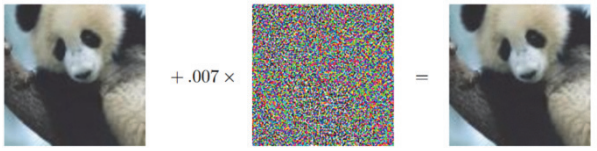


Cognitive instrumentation can support run time monitoring

Provide info for supervisory control

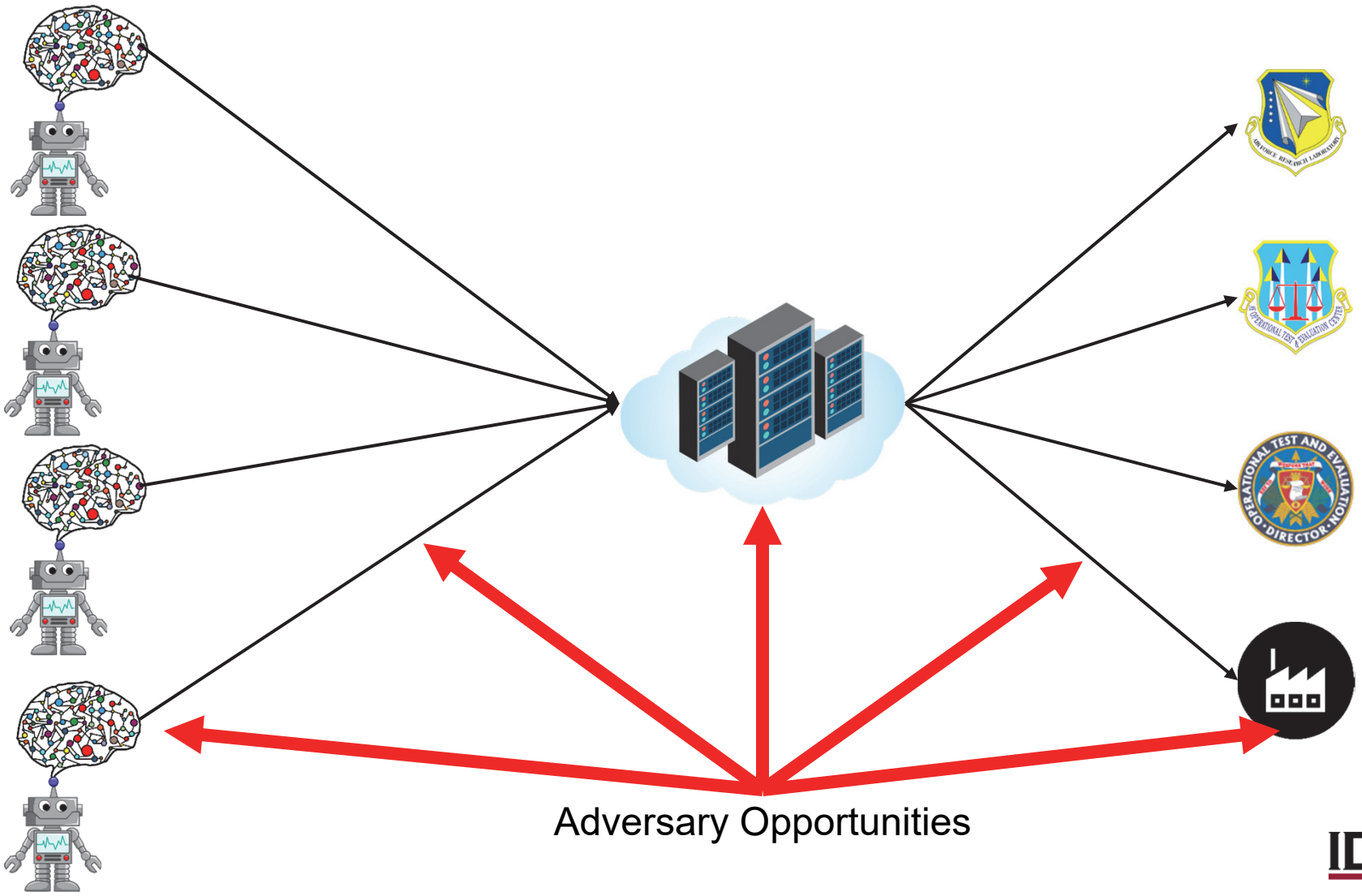
Ensure safe bounds, e.g., TACE

Detect adversarial attacks



Testing of Autonomy in Complex Environments (TACE)

Cognitive instrumentation is only the first step—we need a secure end-to-end data infrastructure too



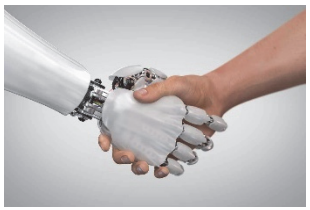
We will need standards and protocols for effective and secure data governance.

Emergent Behaviors

Recommendation: Develop higher-level models to make inferences about agent-to-agent interactive behaviors.

Not all unintended behavior is emergent

- Strong Emergence



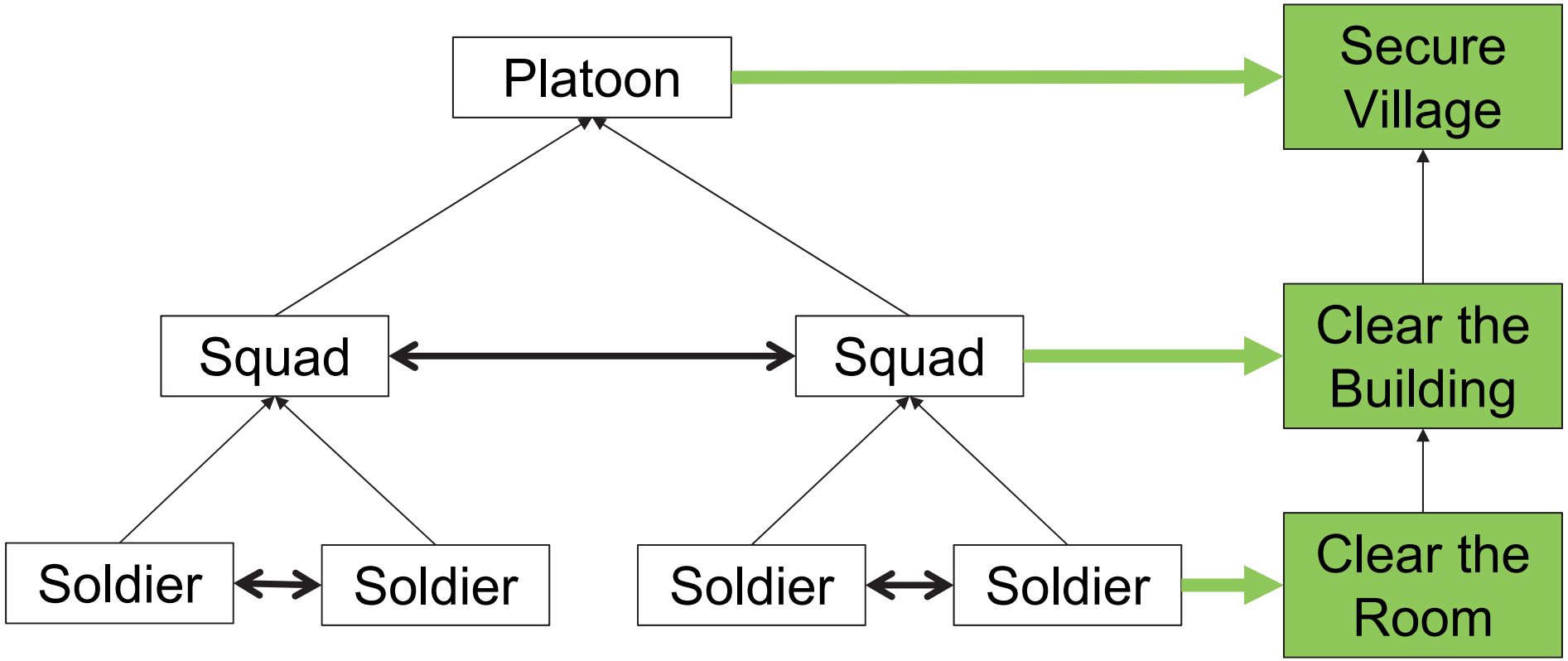
- Weak Emergence

- ~~Bad Design Choices~~

Previous recommendations will help get valid inference for individual systems.

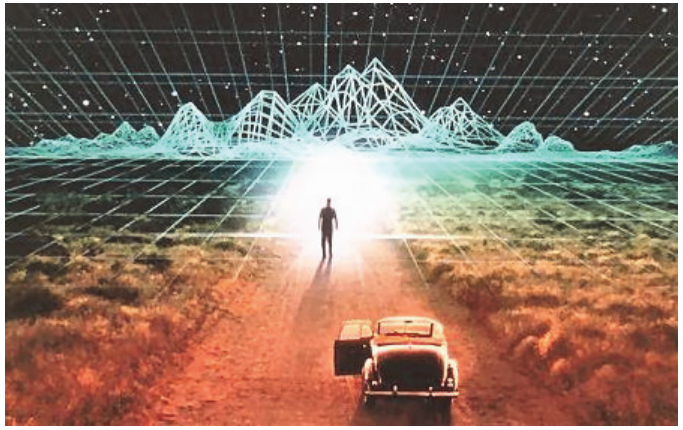
A modified approach is necessary to make models for inter-system behaviors.

Models valid for one level of analysis might not enable inference at a different level



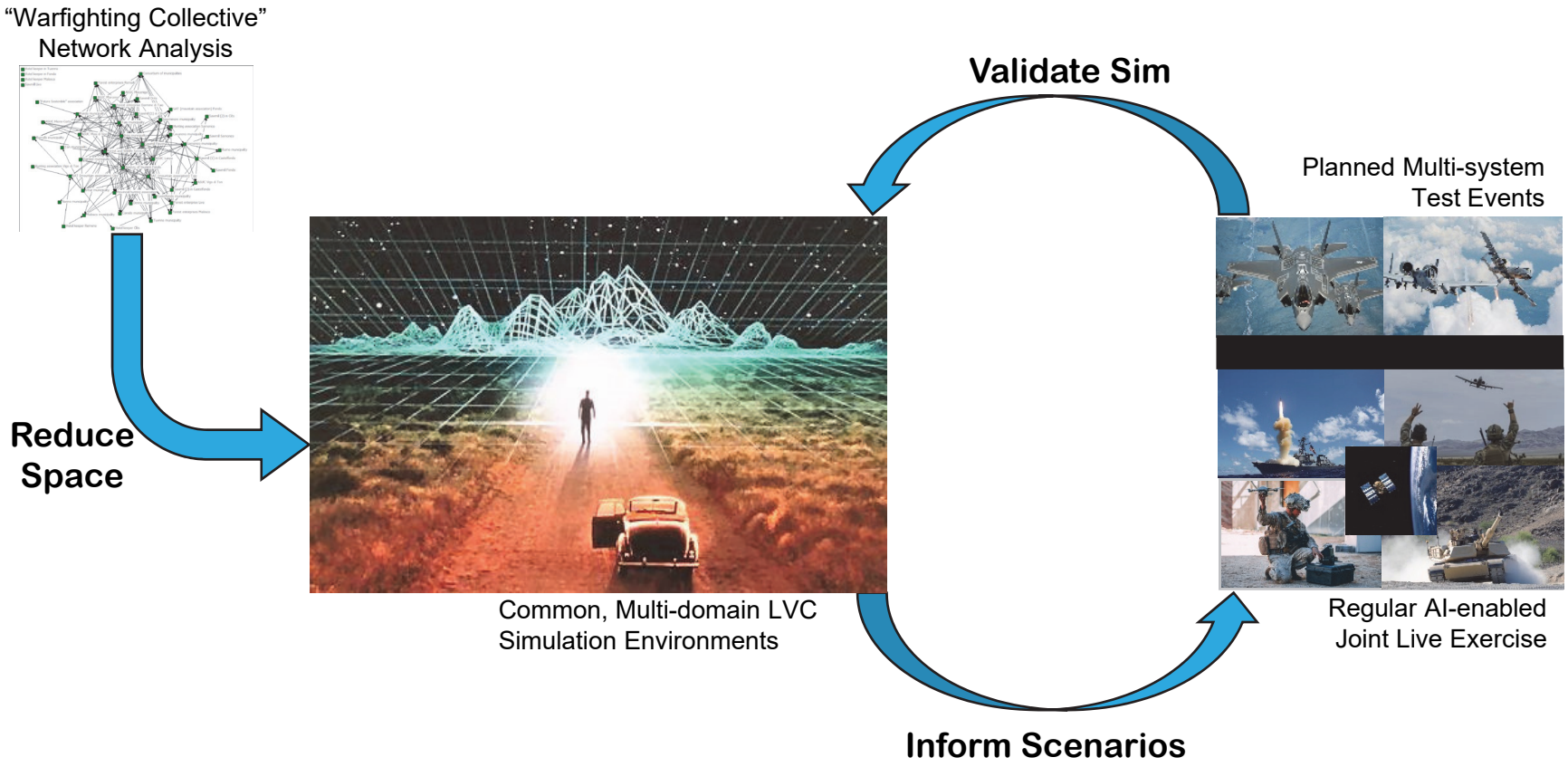
We will need to conduct more explicit testing for agent-agent interactions to look for emergent behaviors

Agent-based Modeling
Multi-agent Simulations



Live Multi-system Testing

Leverage multiple techniques in concert to efficiently cover critical space while validating and diagnosing EB

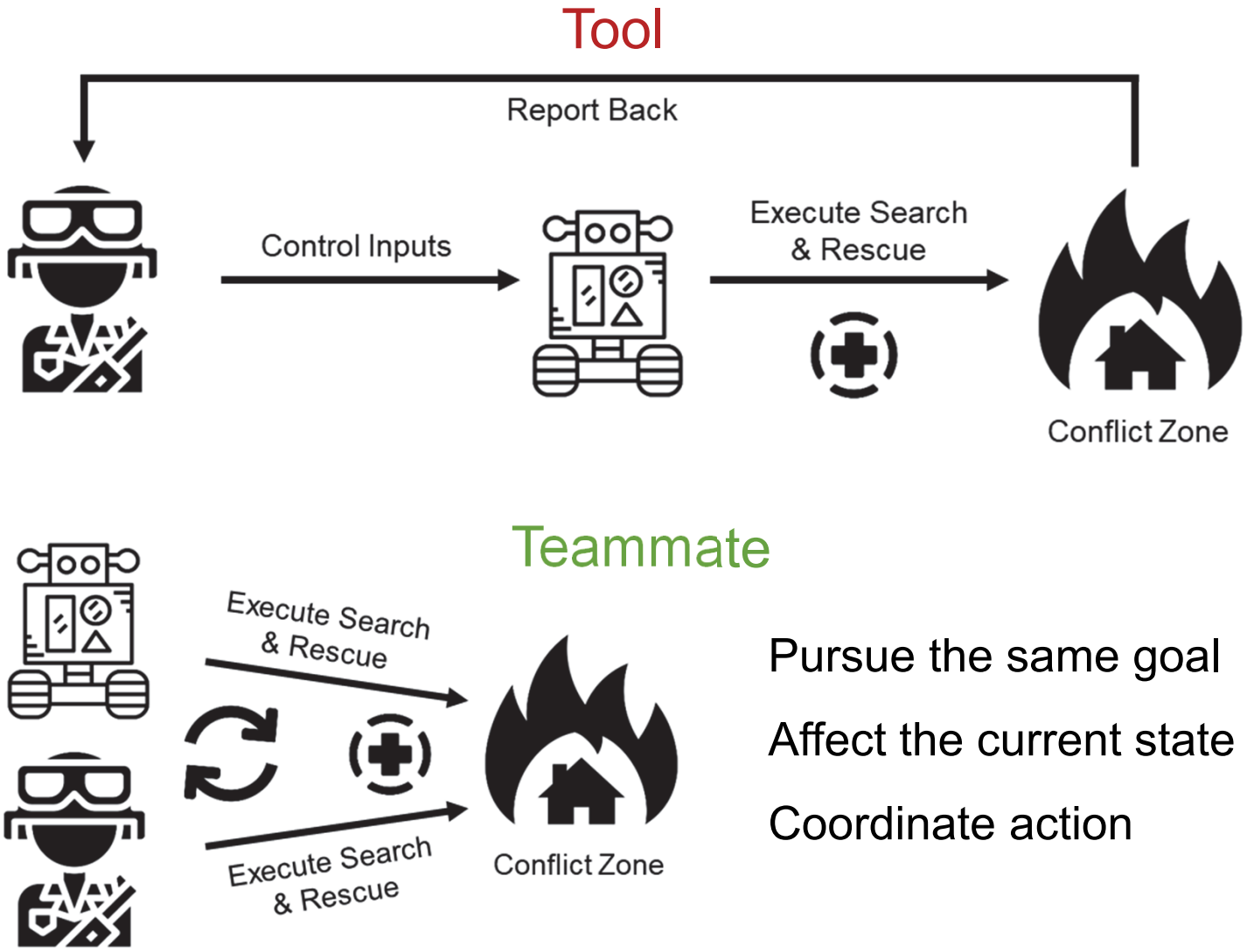


Emergent Behavior (EB); Live, Virtual, Constructive (LCV)

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

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

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

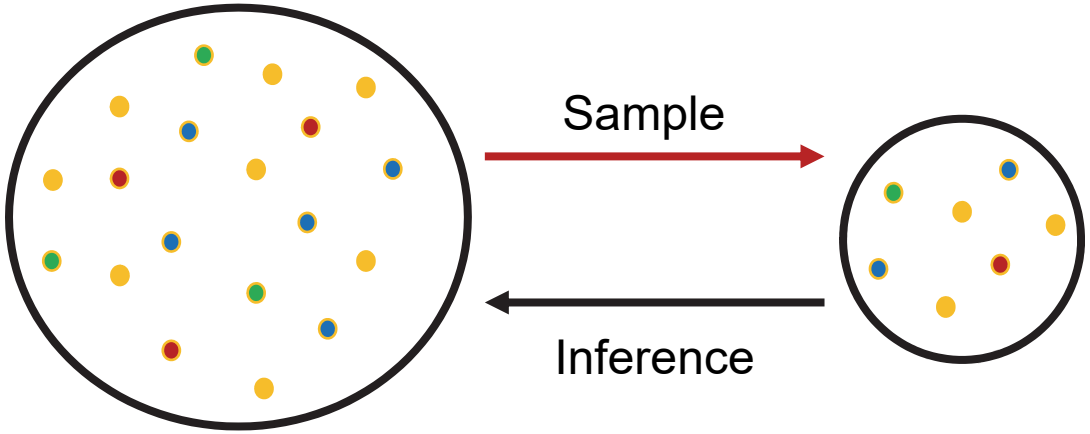


Pursue the same goal

Affect the current state

Coordinate action

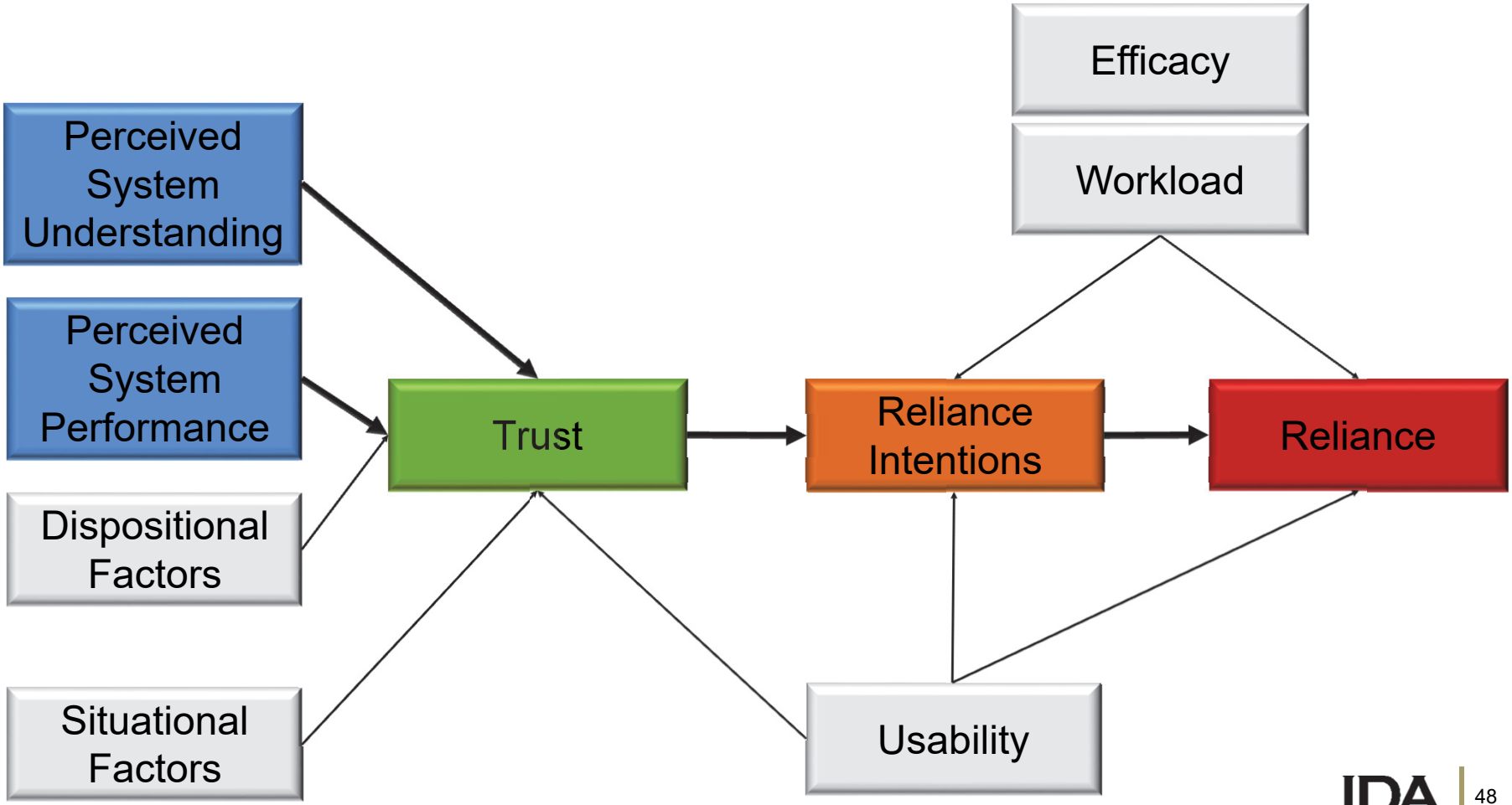
Teammates must be systematically tested with representative groups of operators to permit inference



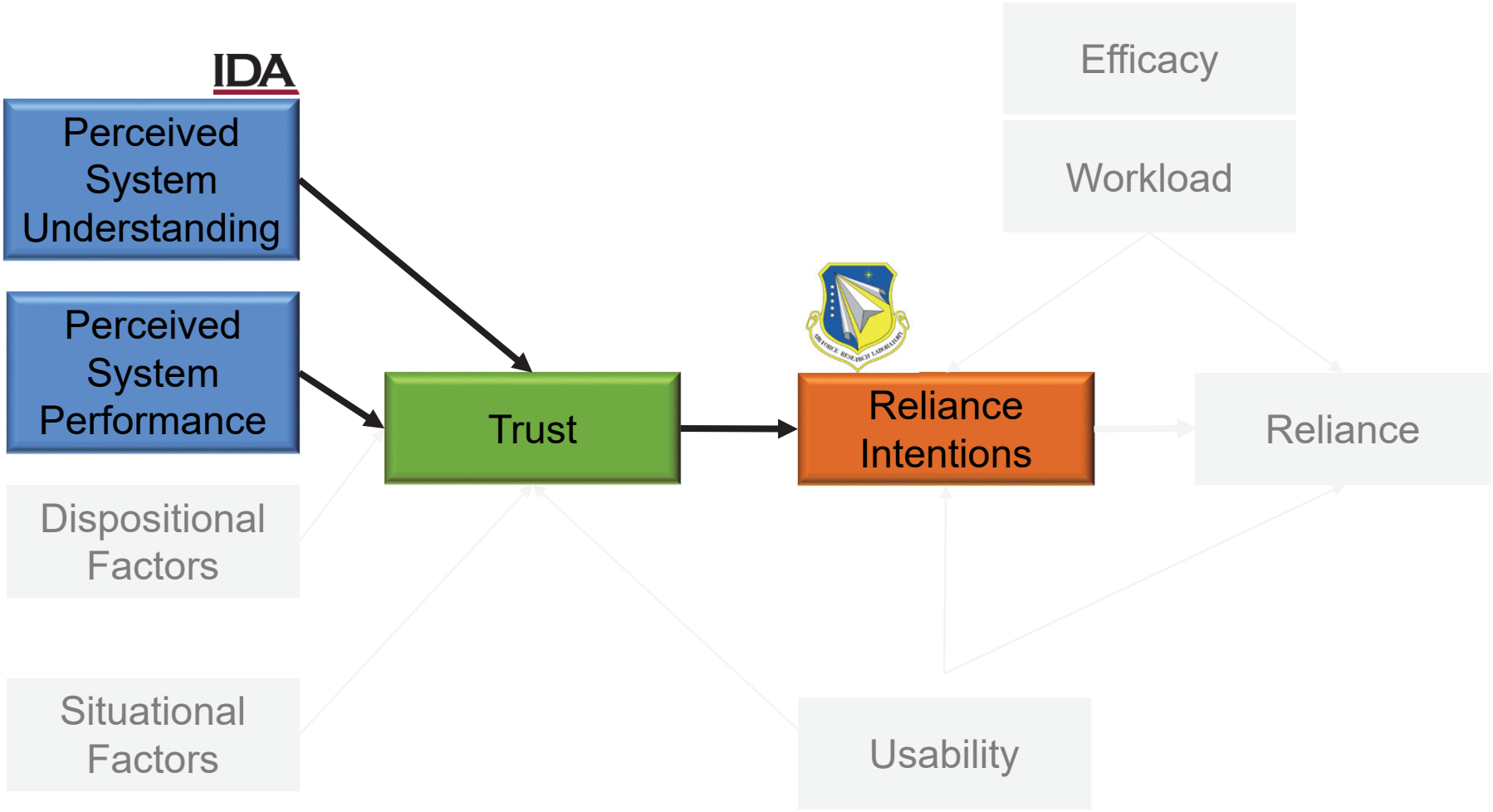
Repeated measure designs are most efficient

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.



IDA validated a measure of contributors to trust; AFRL developed a measure of reliance intentions

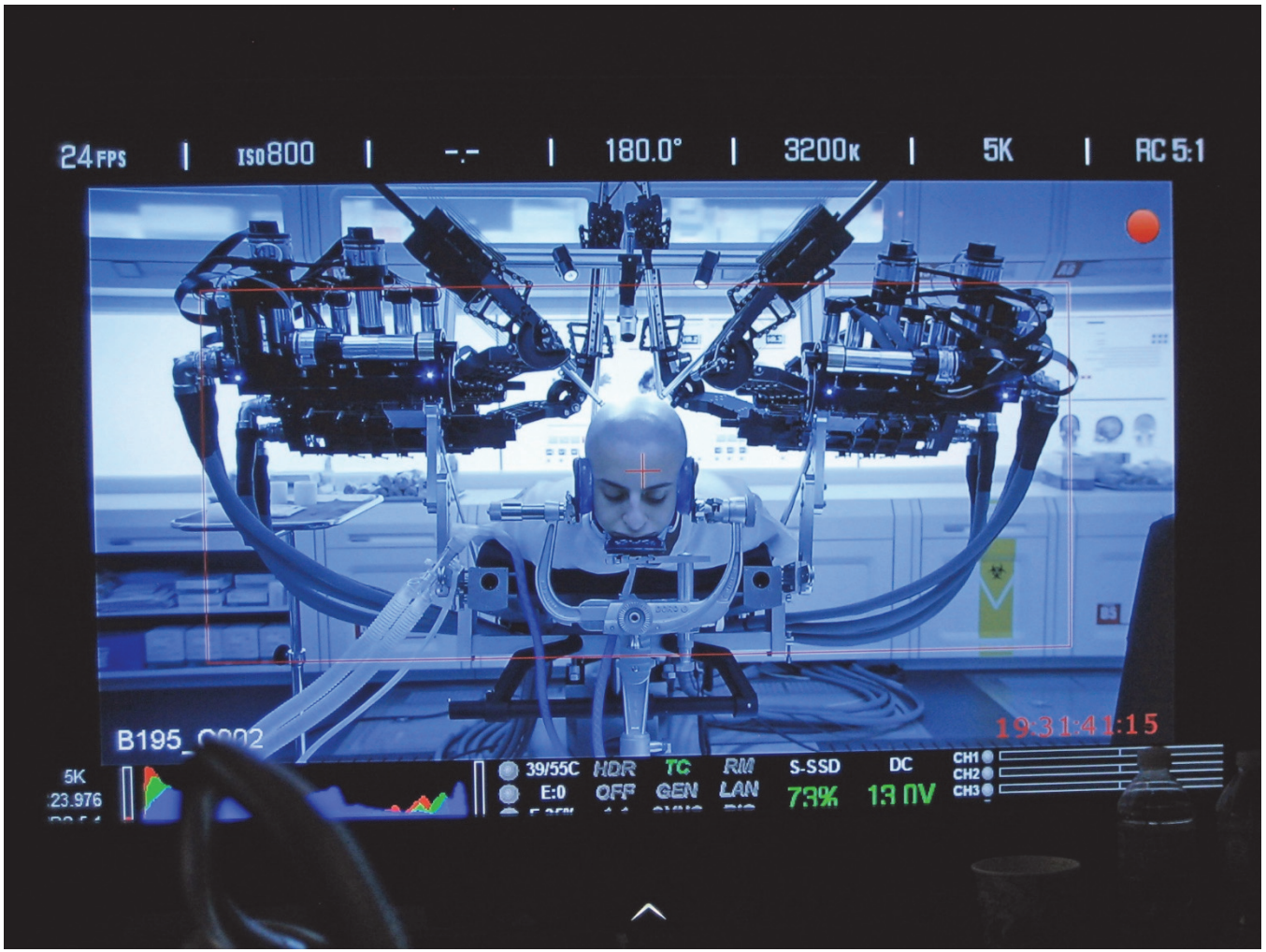


Our team is developing methods and measures to quantify and evaluate the effectiveness of human-machine teams.

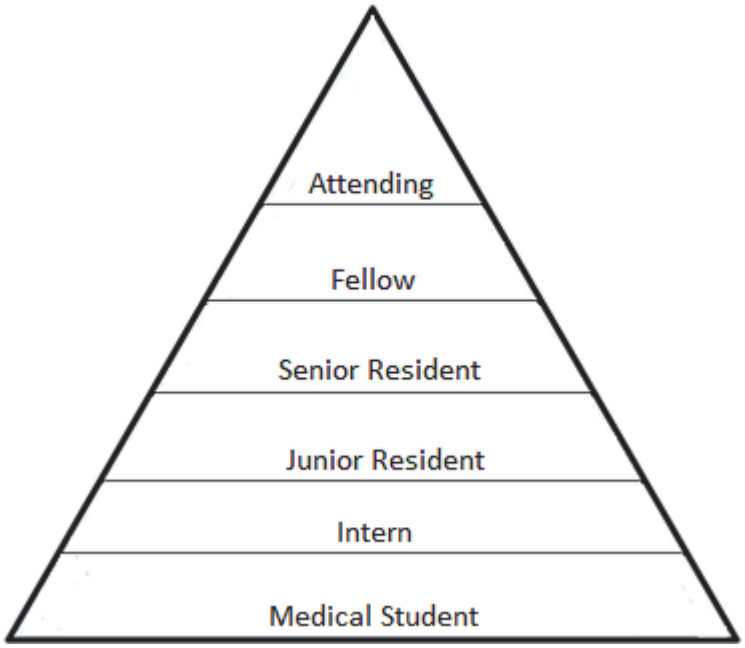
Unintended Behaviors in Evolving Systems

Recommendation: Use human certification methods as the starting but not end point.

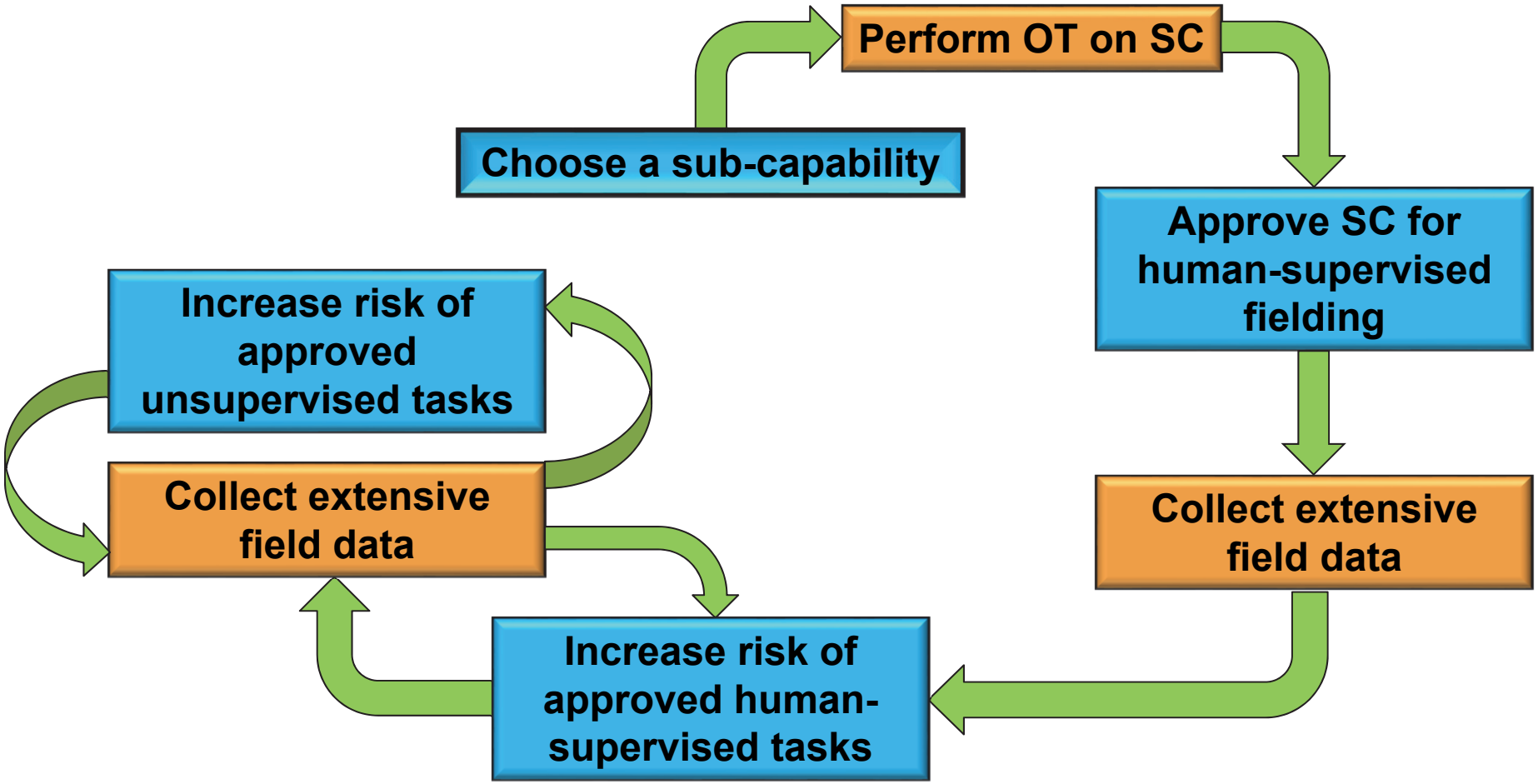
Some systems operate in fundamentally unsafe domains that are difficult to sufficiently simulate



We have the same problem with humans



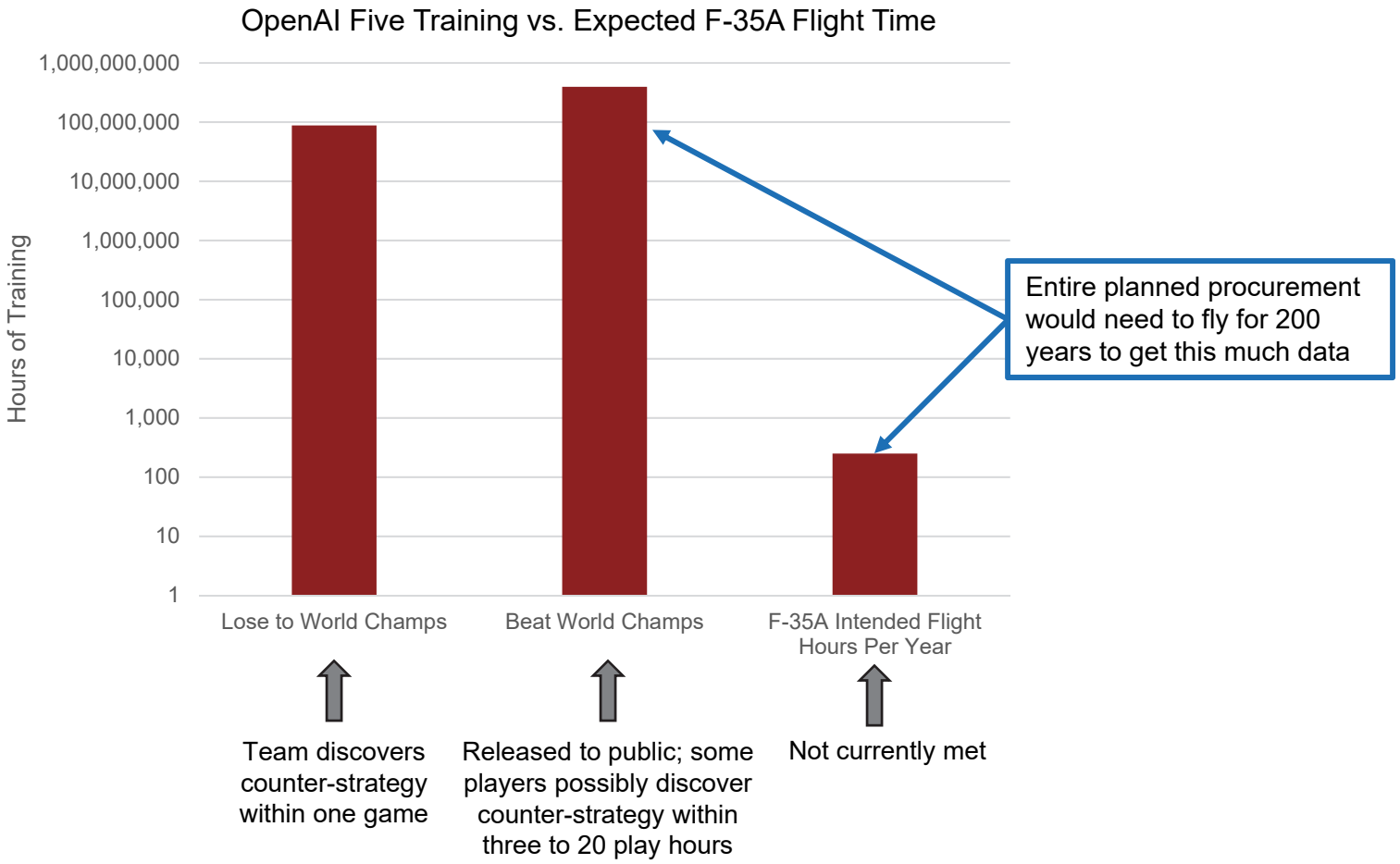
Fundamentally unsafe behaviors can be tested through graded autonomy with limited capability fielding



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



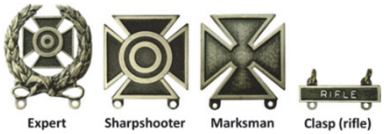
Online learning is really just online change.

We would need to continuously recertify these systems.

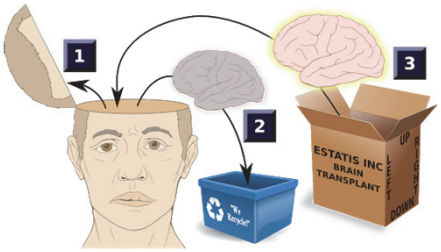
Mitigate model change challenge in online learning through O-I-D levels paradigm for recertification

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.

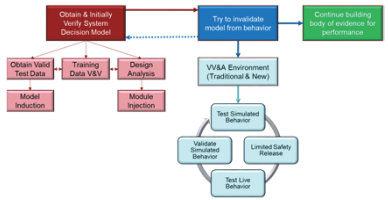


Field Service Representative (FSR); Organizational, Intermediate, Depot (O-I-D)

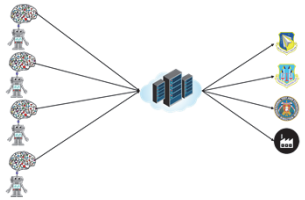
Conclusions

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

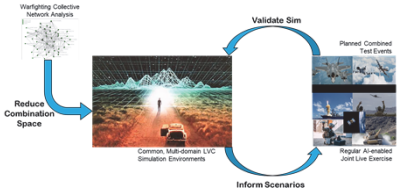
- Obtain, verify, validate, and accredit individual system decision models



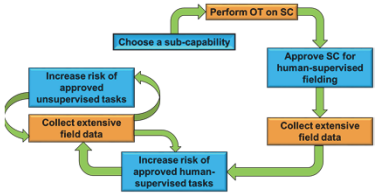
- Diagnose unintended behavior through cognitive instrumentation feeding a secure data infrastructure



- Develop higher-level models to make inferences about inter-agent emergent behaviors



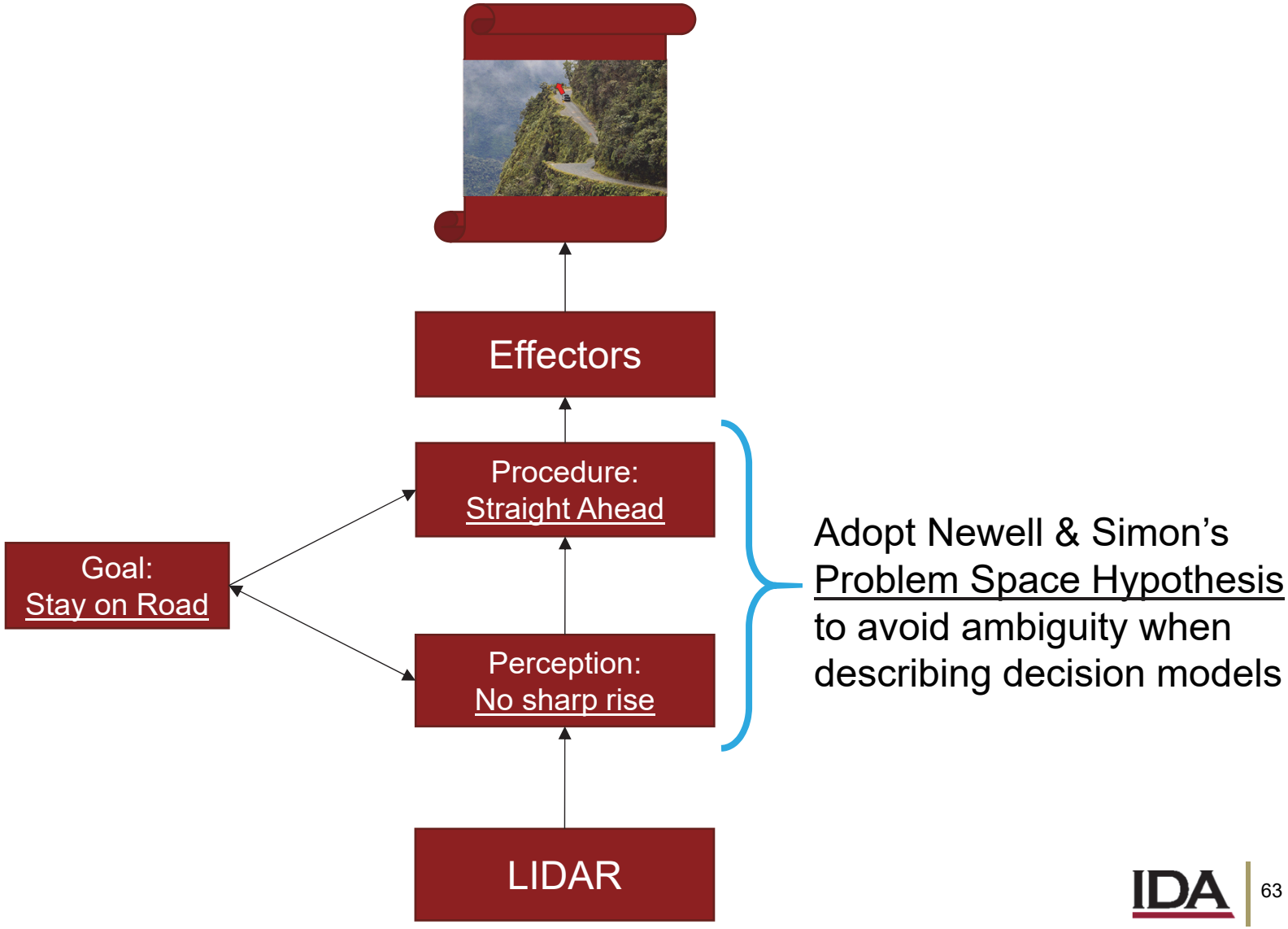
- Evolving systems are probably a far horizon, but human certification methods can be a starting point



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

Backups

Use technical, not colloquial language to describe the components of system decision models



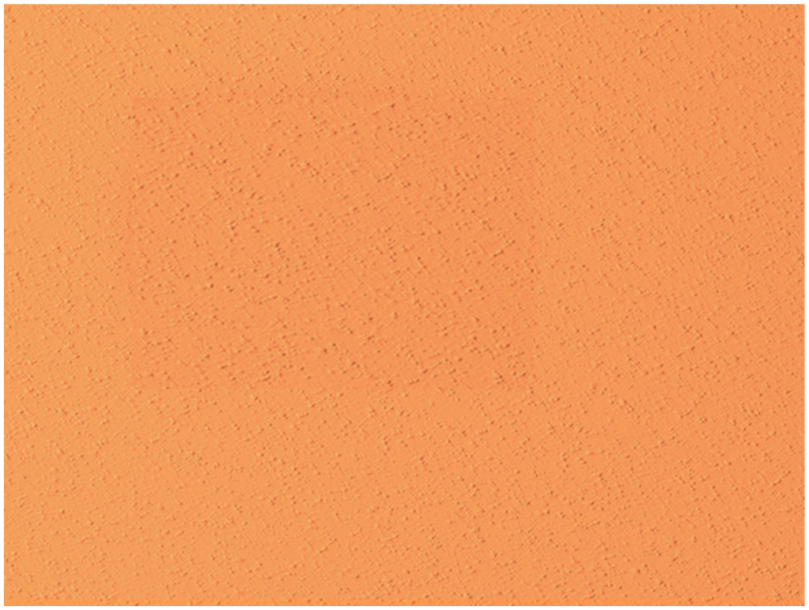
Decision types in the Problem Space Hypothesis will alter what and how we test

- Executive Autonomy: Make decisions about goal states, sub-goals, problem space representations, and path constraints
  - e.g., decision-aides

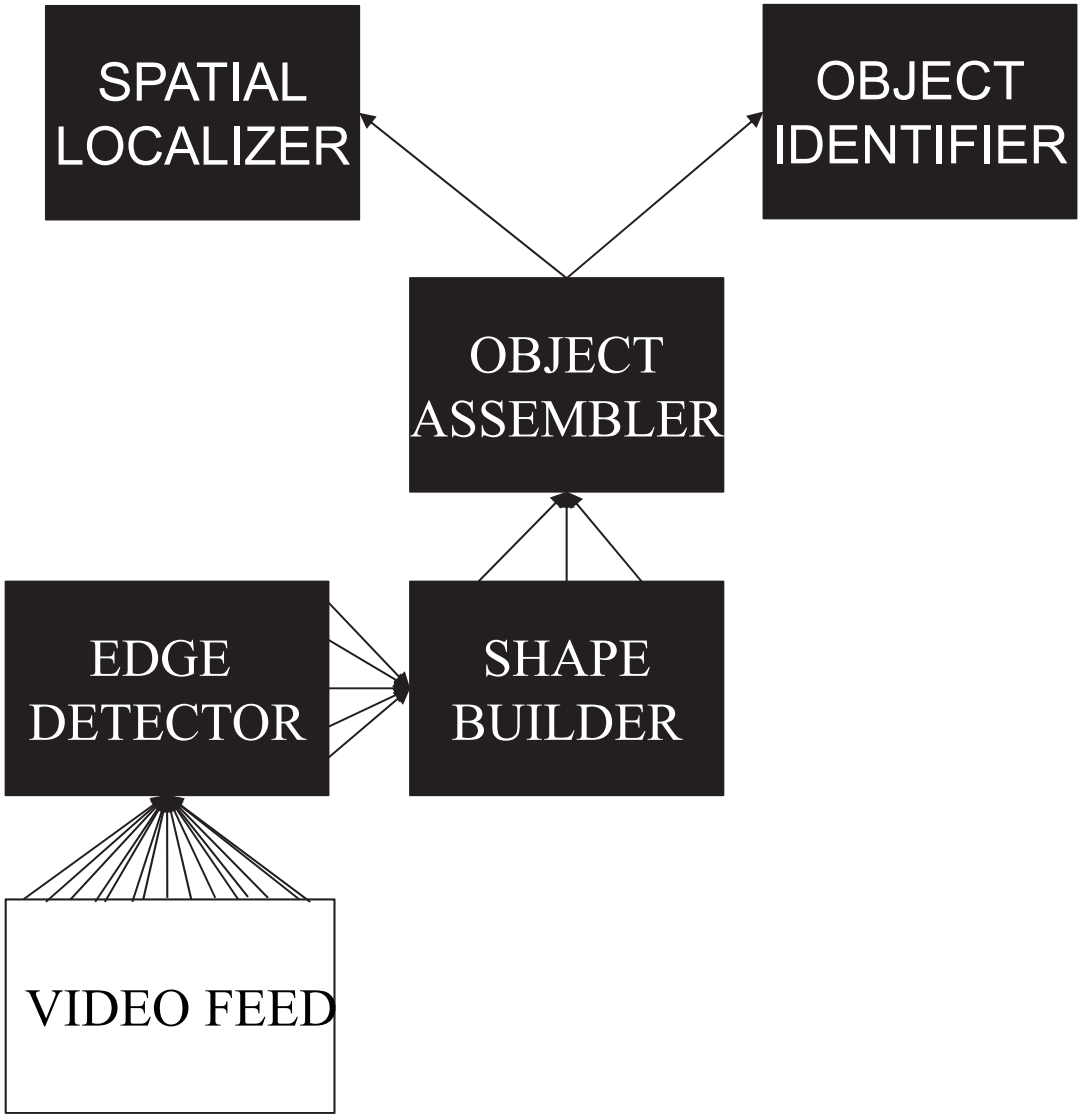
- Perceptual Autonomy: Make decisions about how current problem state is defined
  - e.g., image classifiers

- Procedural Autonomy: Make decisions about next operator/procedure selected
  - e.g., Markov Decision Process

Sensor physics can be valid without the environmental features being valid



Modules example



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02-2020	IDA Publication	

4. TITLE AND SUBTITLE

Briefing to the Air Force Scientific Advisory Board: T&E Contributions to Avoiding Unintended Behaviors in Autonomous Systems

5a. CONTRACT NUMBER

Separate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wojton, Heather
Porter, Daniel

5d. PROJECT NUMBER

5e. TASK NUMBER

C9089

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-12078

H 2020-000049

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

10. SPONSOR/MONITOR'S ACRONYM(S)

IDA

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release.

13. SUPPLEMENTARY NOTES

Project Leader: Heather Wojton

14. ABSTRACT

To provide assurance that AI-enabled systems will behave appropriately across the range of their operating conditions without performing exhaustive testing, the DoD will need to make inferences about system decision making. However, making these inferences validly requires understanding what causally drives system decision-making, which is not possible when systems are black boxes. In this briefing, we discuss the state of the art and gaps in techniques for obtaining, verifying, validating, and accrediting (OVVA) models of system decision-making

15. SUBJECT TERMS

artificial intelligence (AI); autonomous systems; autonomy framework; 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			Heather Wojton (OED)
Unclassified	Unclassified	Unclassified	Unlimited	75	19b. TELEPHONE NUMBER *(Include area code)* (703)845-6811