Neuro-Symbolic Reasoning for Assured Autonomy

Andrzej Banaszuk, Director of Strategy

Alberto Speranzon, Chief Scientist, Autonomy

Mauricio Castillo-Effen, Fellow, Trustworthy AI & Autonomy

Advanced Technology Laboratories, Lockheed Martin

AI Assurance Workshop, August 10, 2023

Outline

- Challenge: Assured and Trusted Collaborative Autonomy
- Sample of ATL assurance approaches
- Challenge Problems in Developing a Neuro-Symbolic OODA Loop
- Directions for future research

ATL Strategic Initiatives

Autonomy and Crewed-Uncrewed Teaming

- Human-Autonomy Collaborative Teams
- Cognitive Load Optimization
- Assurance and Trust

Neuro-Symbolic Reasoning

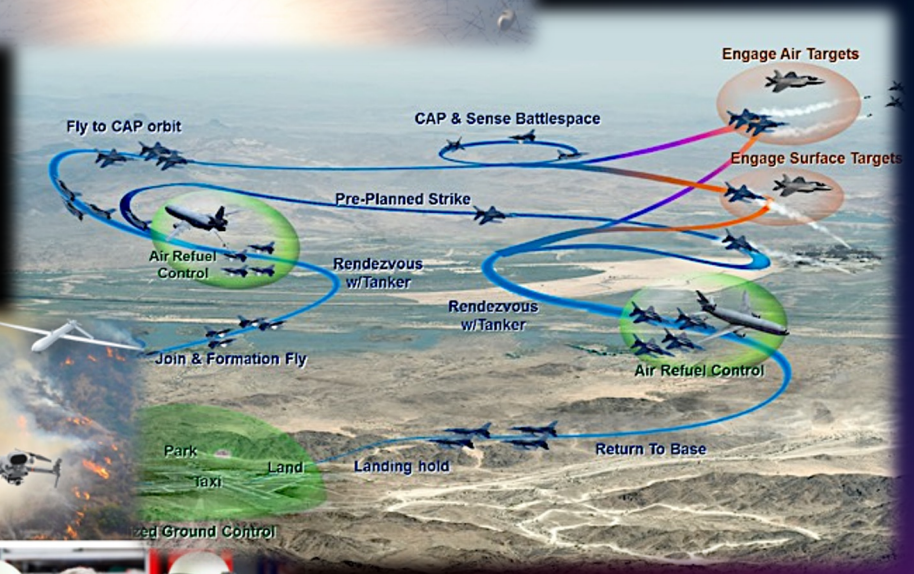
- Uncertainty and Ambiguity Management
- Causal Generative Models
- Lifelong Learning at the Edge

Neuro-Symbolic Reasoning

- Scalable Reinforcement Learning
- Composable AI
- Generalizable AI

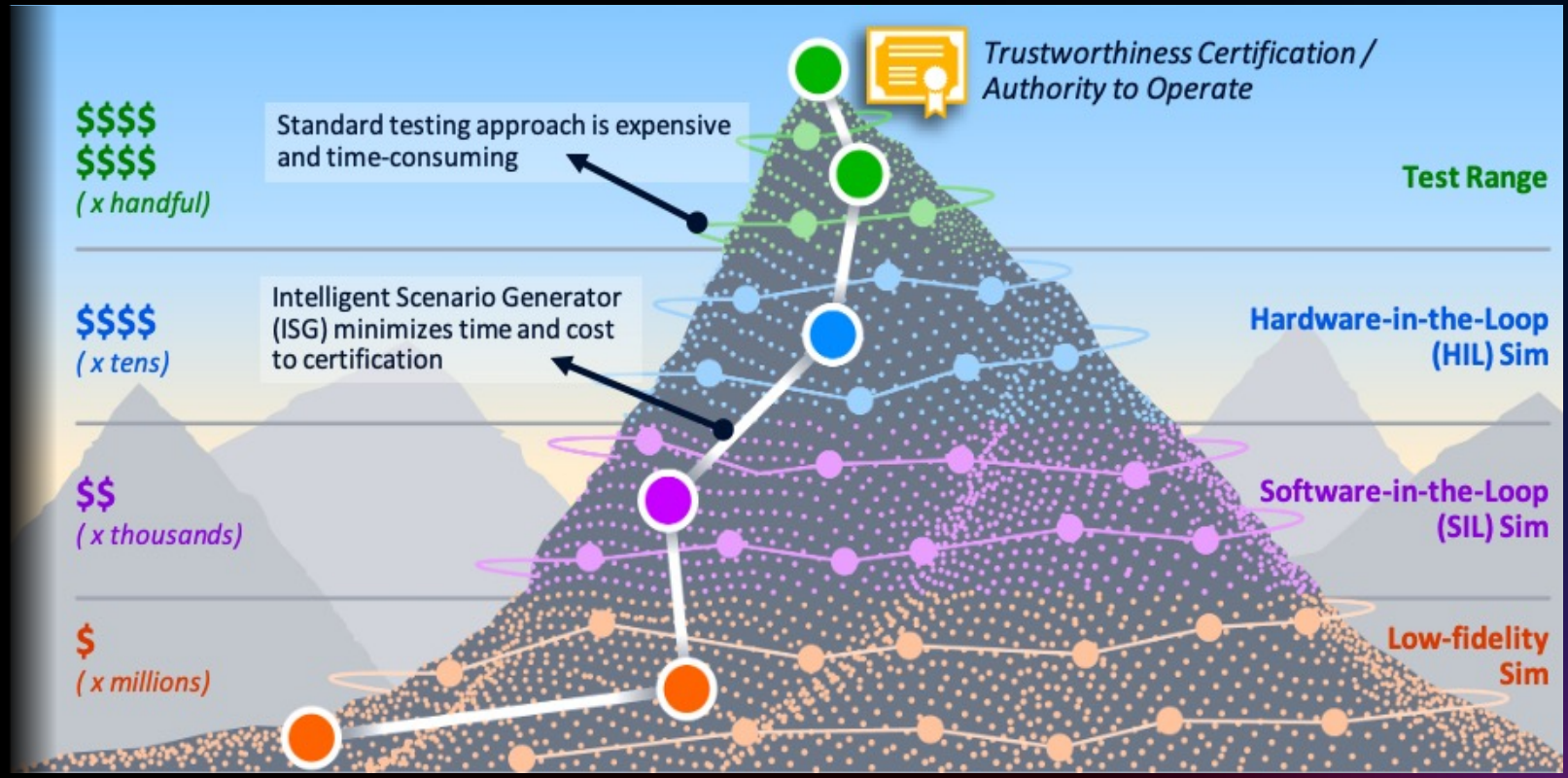
Challenge: Assured and Trusted Collaborative Autonomy

- Human-autonomy collaborative teams
- Multi-scale and multi-domain
- Scalability and composability
- Uncertain and ambiguous environments



ATL Assurance Research: Scenario Generation

- Defines ***what and how to test***
- Three categories of experiments:
  - Targeted discovery (“known unknowns”)
  - Exploratory discovery (“unknown unknowns”)
  - Evidence gathering (“justified confidence”)
- Technologies:
  - Ability to express and reason about scenarios as distributions
  - Generation of information-lucrative experiments
  - Architecture for scalability
  - Software-enabled Simulation-Emulation-Stimulation

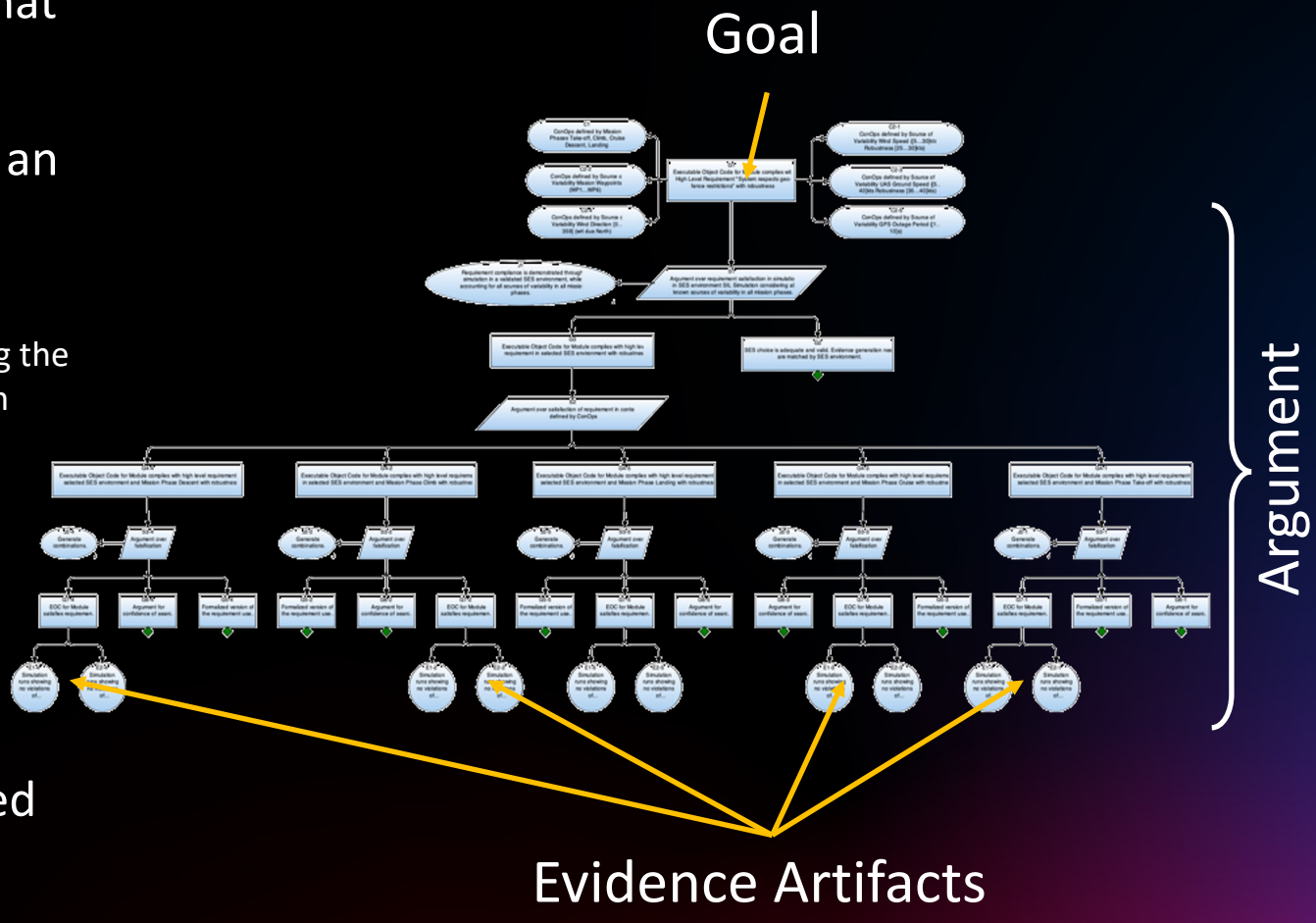


ATL Assurance Research: Assurance Cases

- A means of increasing well-founded confidence that a system will behave as intended
- An assurance case requires claims, evidence, and an argument linking evidence to claims:

  Evidence
  > Results of observing, analyzing, testing, simulating, and estimating the properties of a system that provide fundamental information from which the presence of some system property can be inferred

  High Level Argument
  > Explanation of how the available evidence can be reasonably interpreted as indicating acceptable operation, usually by demonstrating compliance with requirements, sufficient mitigation of hazards, avoidance of hazards, etc.



- Claims with no supporting evidence are unfounded
- Evidence without claims is unexplained
- Argument-based approaches: UL4600 “Standard for Safety for the Evaluation of Autonomous Vehicles and Other Products”

Neuro-Symbolic Reasoning: Motivation

Challenges addressed

- Scalability and composability
- Human-autonomy collaborative teams

Deep Neural Networks

- Large training data sets required
- Difficult to incorporate physics, human expert knowledge
- Limited applicability outside of learning data sets
- Susceptible to adversarial attacks
- Simulation-based Reinforcement Learning not scalable
- Limited composability
- Limited explainability
- **Limited verification approaches**

E. Cohen, Y.Y. Elboher, C. Barrett, and G. Katz, Tighter Abstract Queries in Neural Network Verification, EPiC Series in Computing , Vol. 94, 2023.

H. Hanspaal, A. Lomuscio, Efficient Verification of Neural Networks LVM-based Specifications, Proceedings of the 36th IEEE CVPR23.

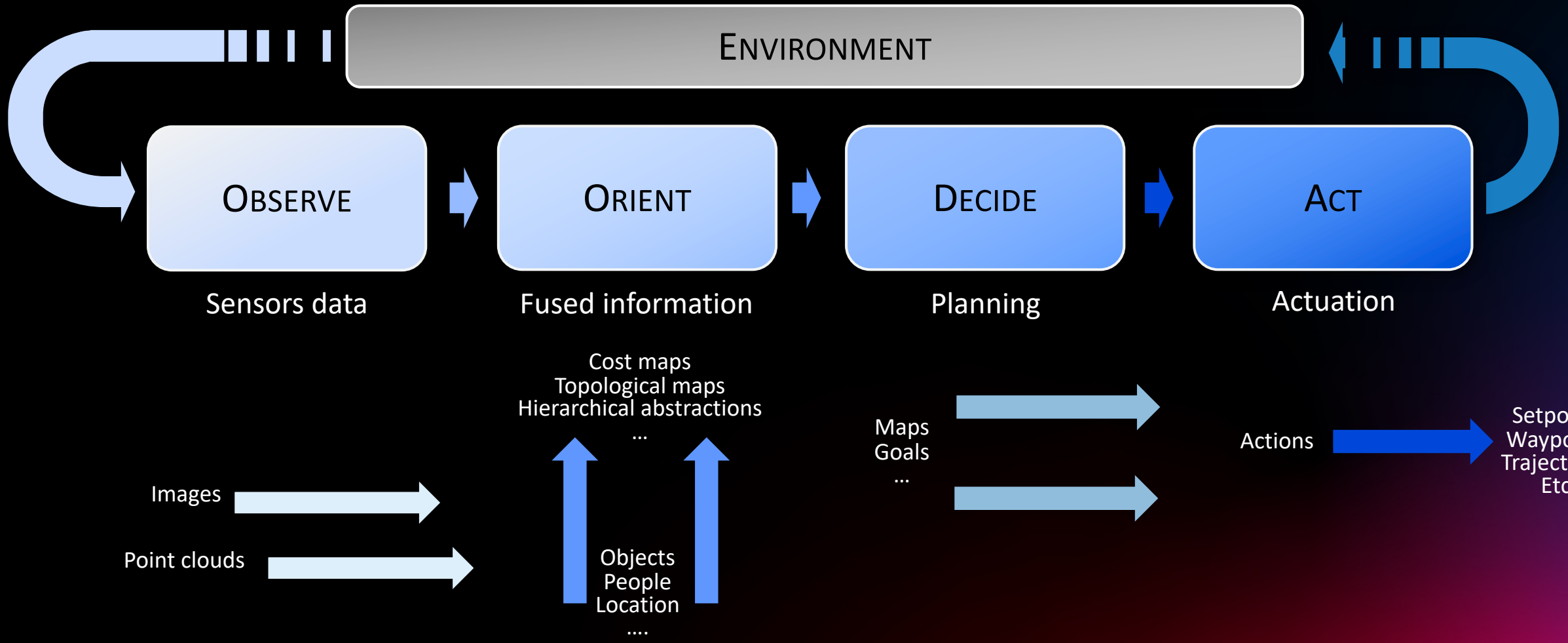
Neuro-Symbolic Reasoning

- Reduces data requirements by orders of magnitude
- Enables capturing domain knowledge
- Extends applicability (knowledge captured in symbolic layers)
- Reduces susceptibility (contextual reasoning)
- Accelerates Reinforcement Learning (symbolic abstractions)
- Builds-in composability
- Enables explainability and human dialog
- Enables formal verification (symbolic layers)
- Reduces verification effort (sub-symbolic layers)

A. Speranzon, C.H. Debrunner, D. Rosenbluth, M. Castillo-Effen, A.R. Nowicki, K. Alcedo and A. Banaszuk, Challenge Problems in Developing a Neuro-Symbolic OODA Loop, NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning

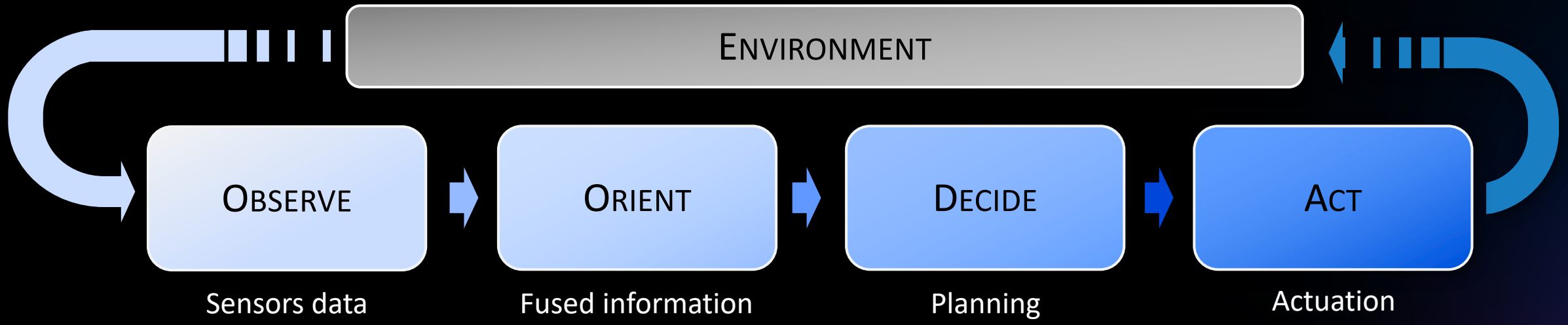
Observe, Orient, Decide and Act (OODA) Loop

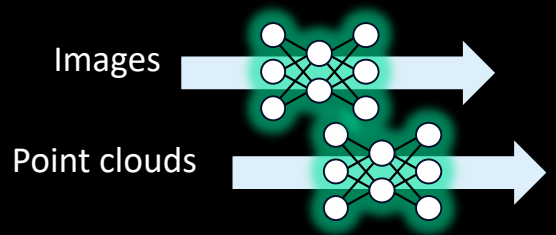
Typical architecture of an autonomous system includes:

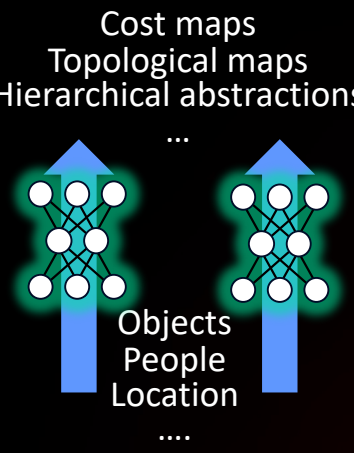


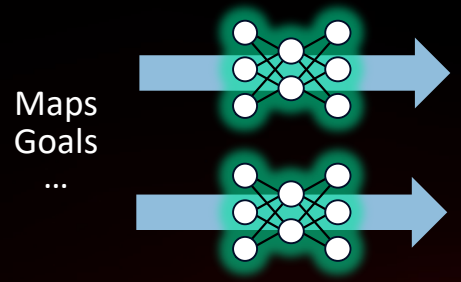
Observe, Orient, Decide and Act (OODA) Loop

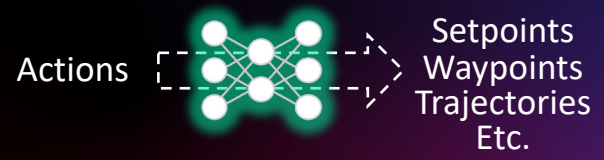
Typical architecture of an autonomous system includes:







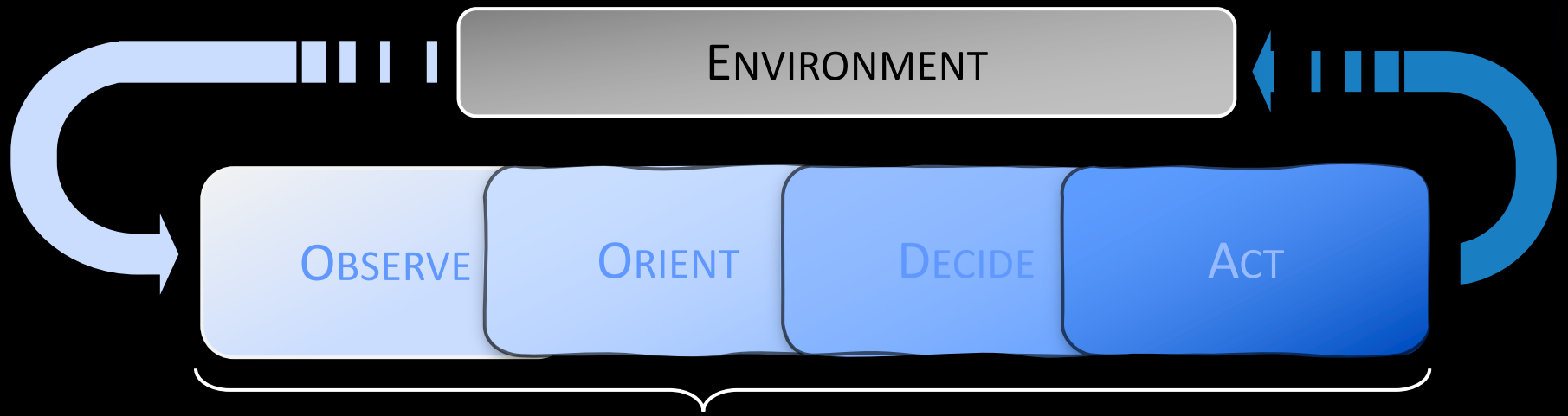


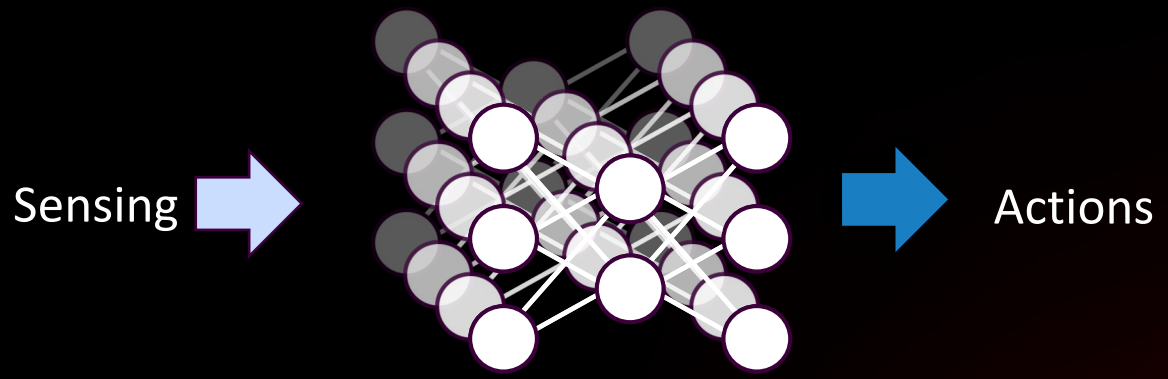


Deep Networks perform many *key processing steps* within the subfunctions of OODA loop

A “Different” Observe, Orient, Decide and Act (OODA) Loop

One can conceive decision loops where representations are (mostly) *hidden* within the layers of DNNs:

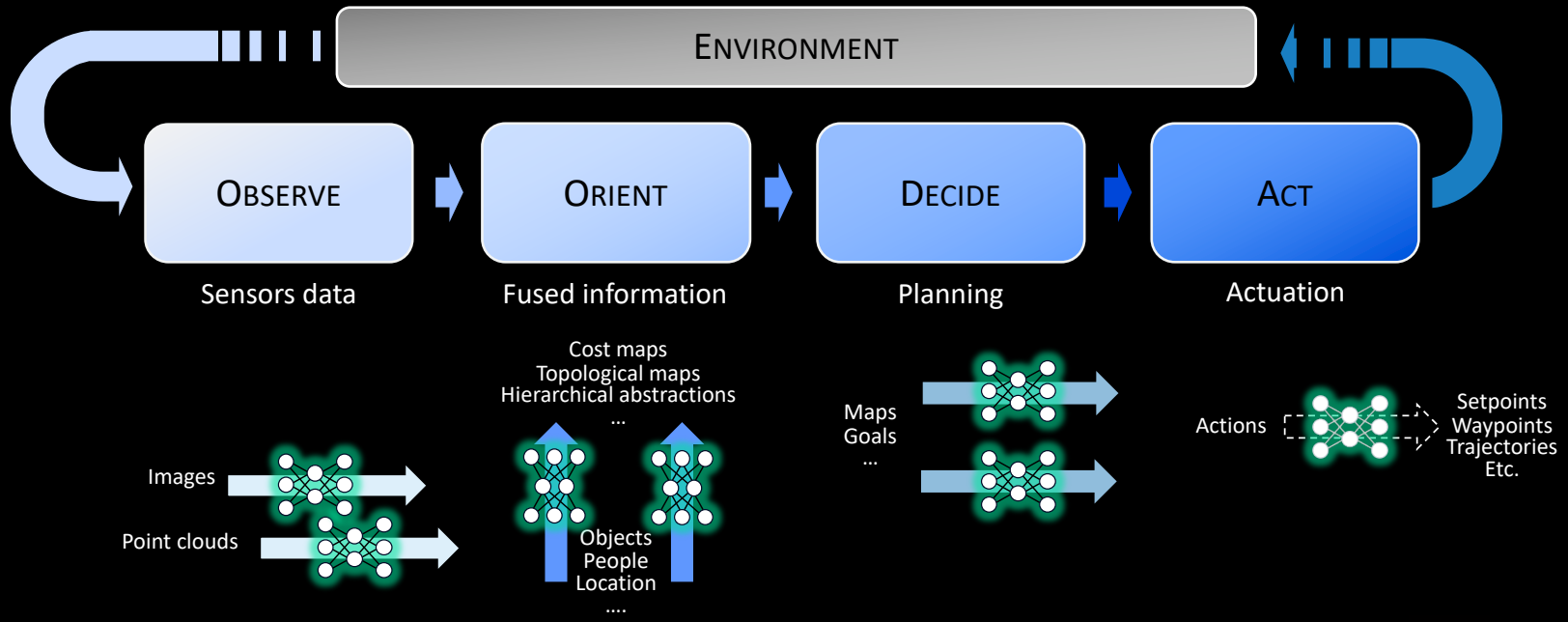




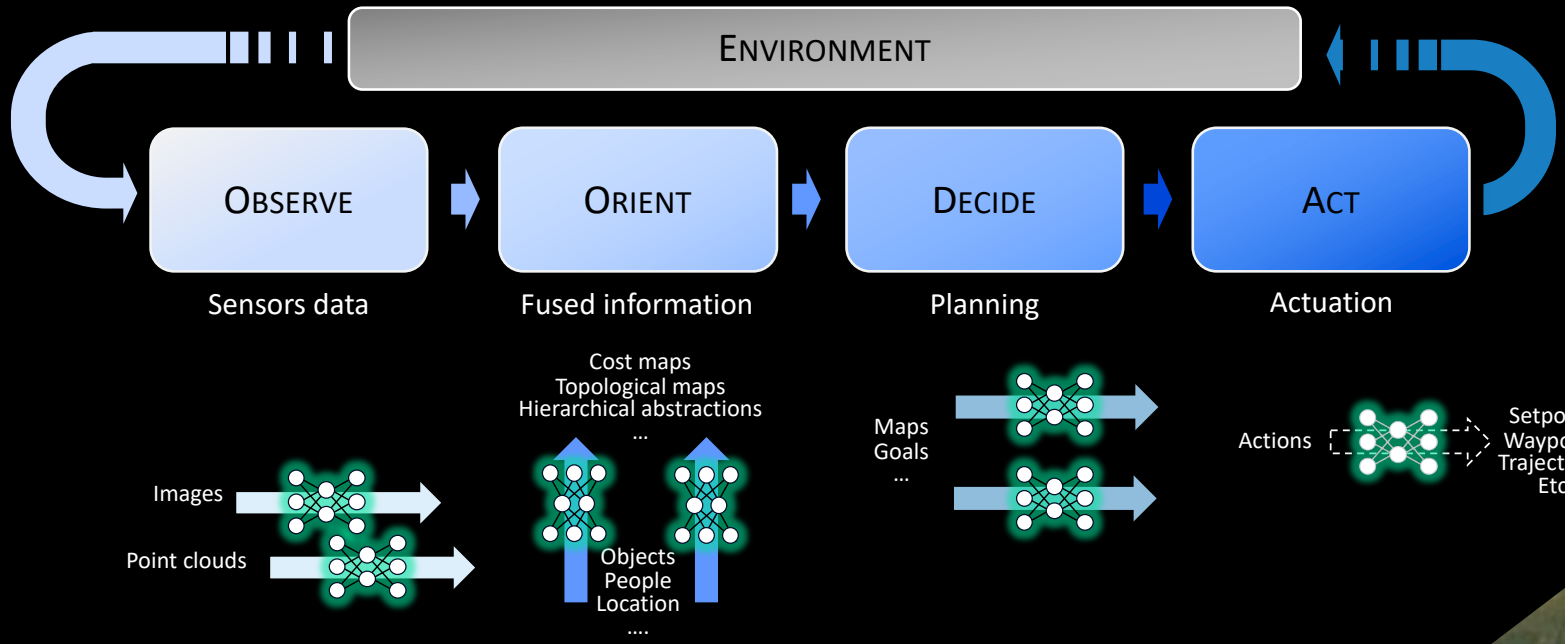
However:

- Difficult to verify and validate especially in safety critical settings
- Data may not be sufficient and/or available
- Simulations too slow to be run and/or too complex/costly to be built
- Difficult to provide operators interpretable insights on what led to certain decisions

Observe, Orient, Decide and Act (OODA) Loop in This Talk



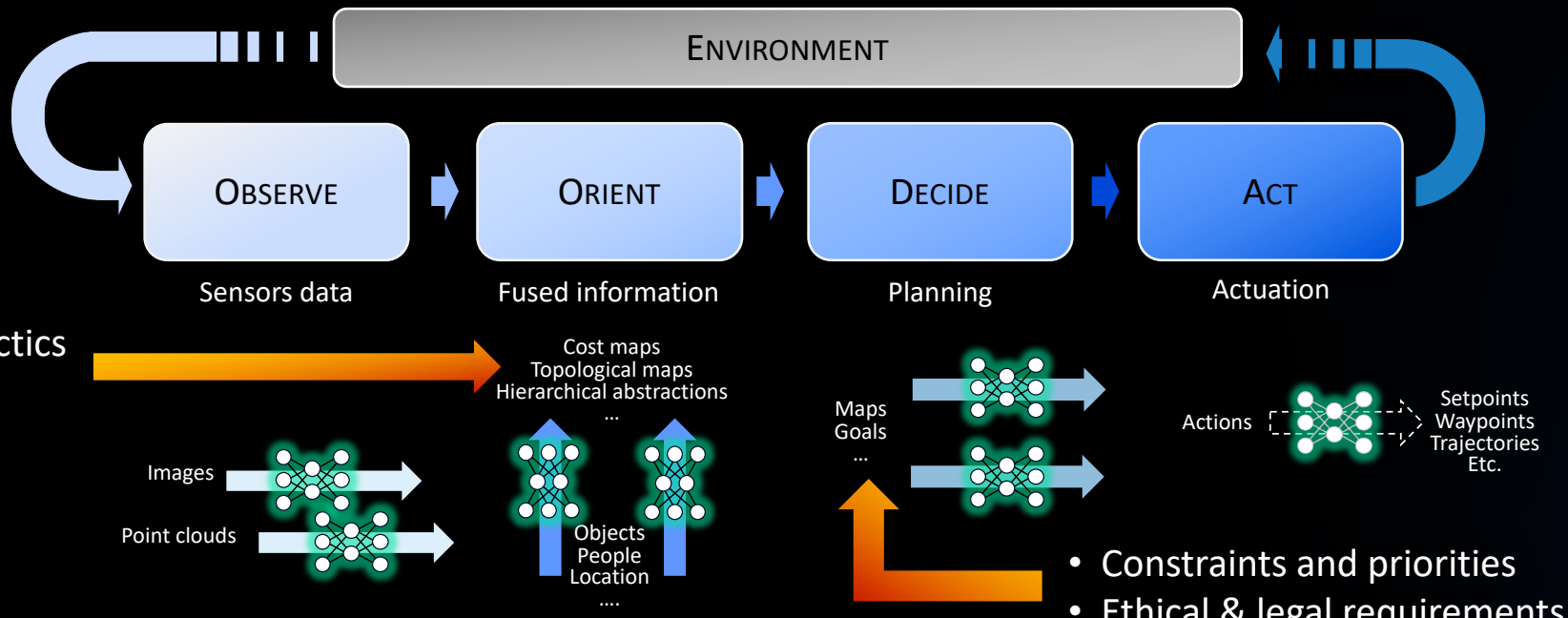
Observe, Orient, Decide and Act (OODA) Loop in This Talk



https://blogs.nvidia.com/blog/2021/11/09/lockheed-martin-wildfires-ai/

- Autonomous platforms with multiple sensors detecting fire fronts and predicting their propagation
- Coordination between humans and autonomous systems requires following specific guidelines
- Decisions on where to drop retardants or water, where to light controlled fires or remove burnable material

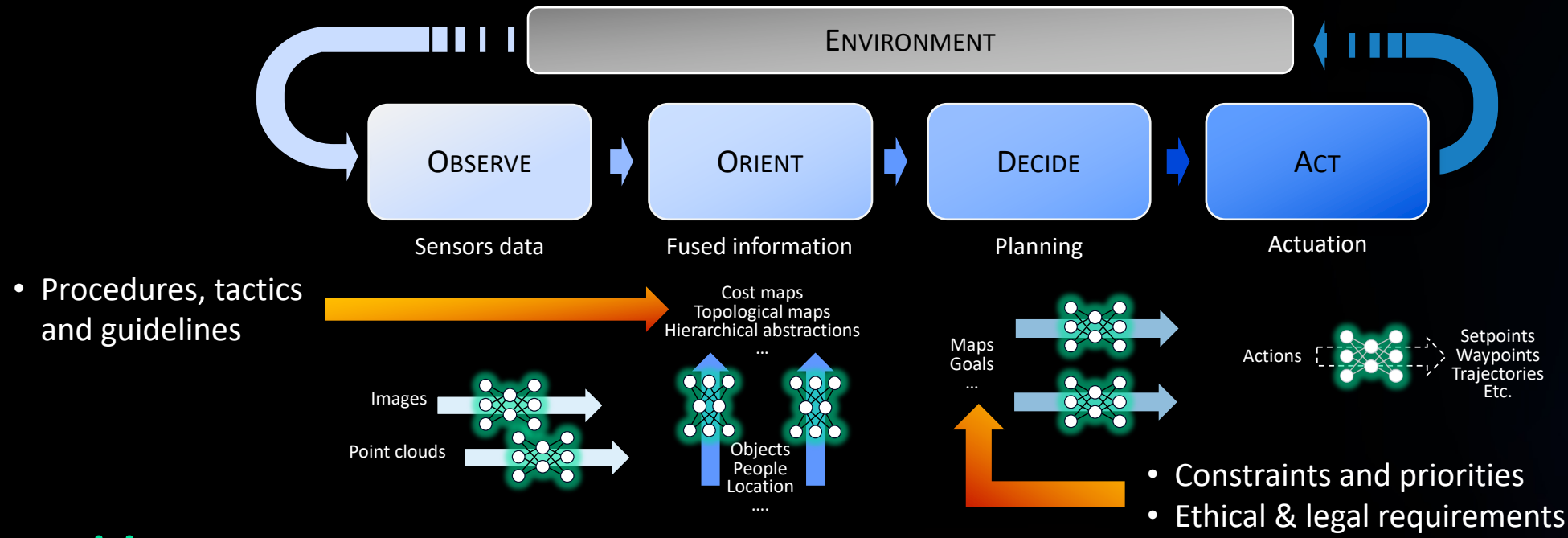
Integration of Prior Knowledge within the OODA Loop



Challenges:

- Data versus symbolic knowledge -- “more/less” data versus “less/more” axioms dichotomy
- Should symbolic knowledge capture be adapted/re-defined for NS architectures?
  - Knowledge capture is typically done to share common understanding, for reusability, make domain assumptions explicit, separate domain from operational knowledge, analyze/query domain knowledge;
  - In various NS approaches knowledge is typically very focused and used to shape constraint/loss functions
- Should symbolic priors be represented/integrated differently in eachpart of the OODA loop?

Integration of Prior Knowledge within the OODA Loop

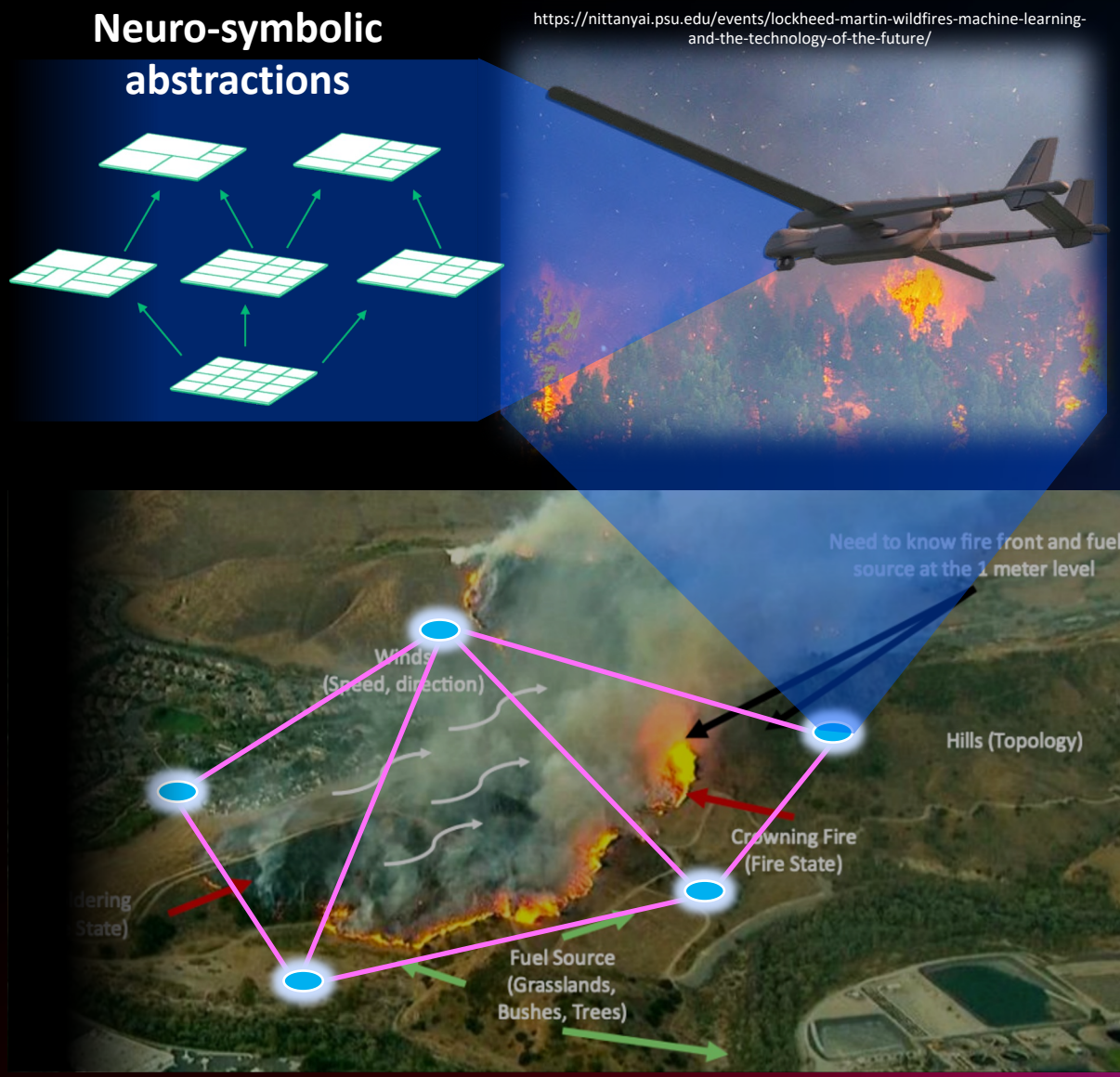


Opportunities:

- Multi-Modal Large Language Models (mmLLMs) may provide an approach to extract knowledge in a semi-automatic fashion, reducing the cost of the task
- ***However***, query engineering to extract meaningful/useable ontologies may be, at this stage, as complex as developing an ontology itself and need to mitigate hallucination, model refinement, etc.

Development of NS Abstractions

- In “Orient” and “Decide” of the OODA loop an autonomous system *generates abstractions* to represent the world and plan relevant decisions
- *Symbolic representations* provide a very appealing way to model such abstractions

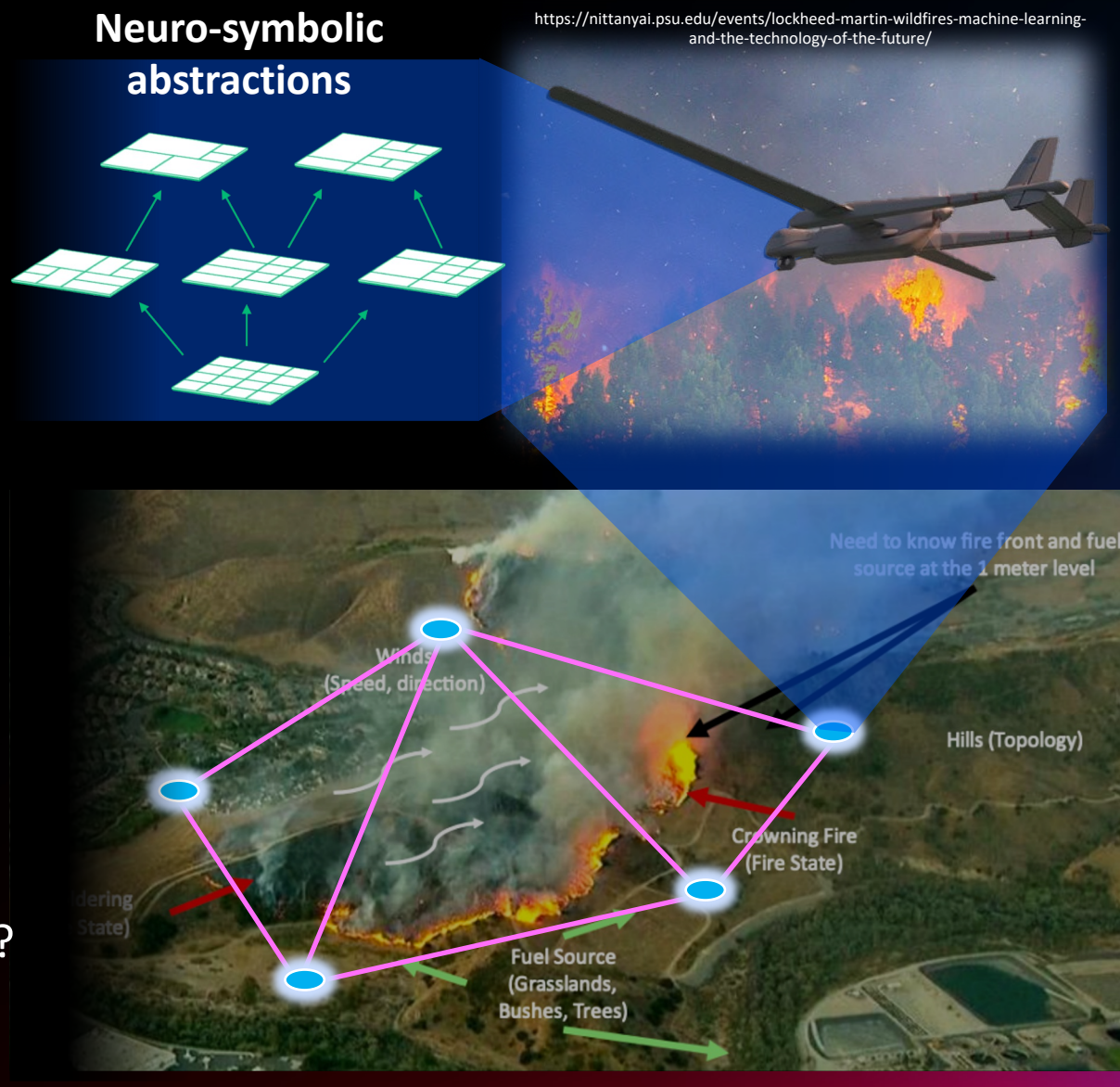


Development of NS Abstractions

- In “Orient” and “Decide” of the OODA loop an autonomous system *generates abstractions* to represent the world and plan relevant decisions
- *Symbolic representations* provide a very appealing way to model such abstractions
  - ❖ Compositionality
  - ❖ Reusability & Transferability
  - ❖ Interpretability
  - ❖ Complexity reduction/Compression

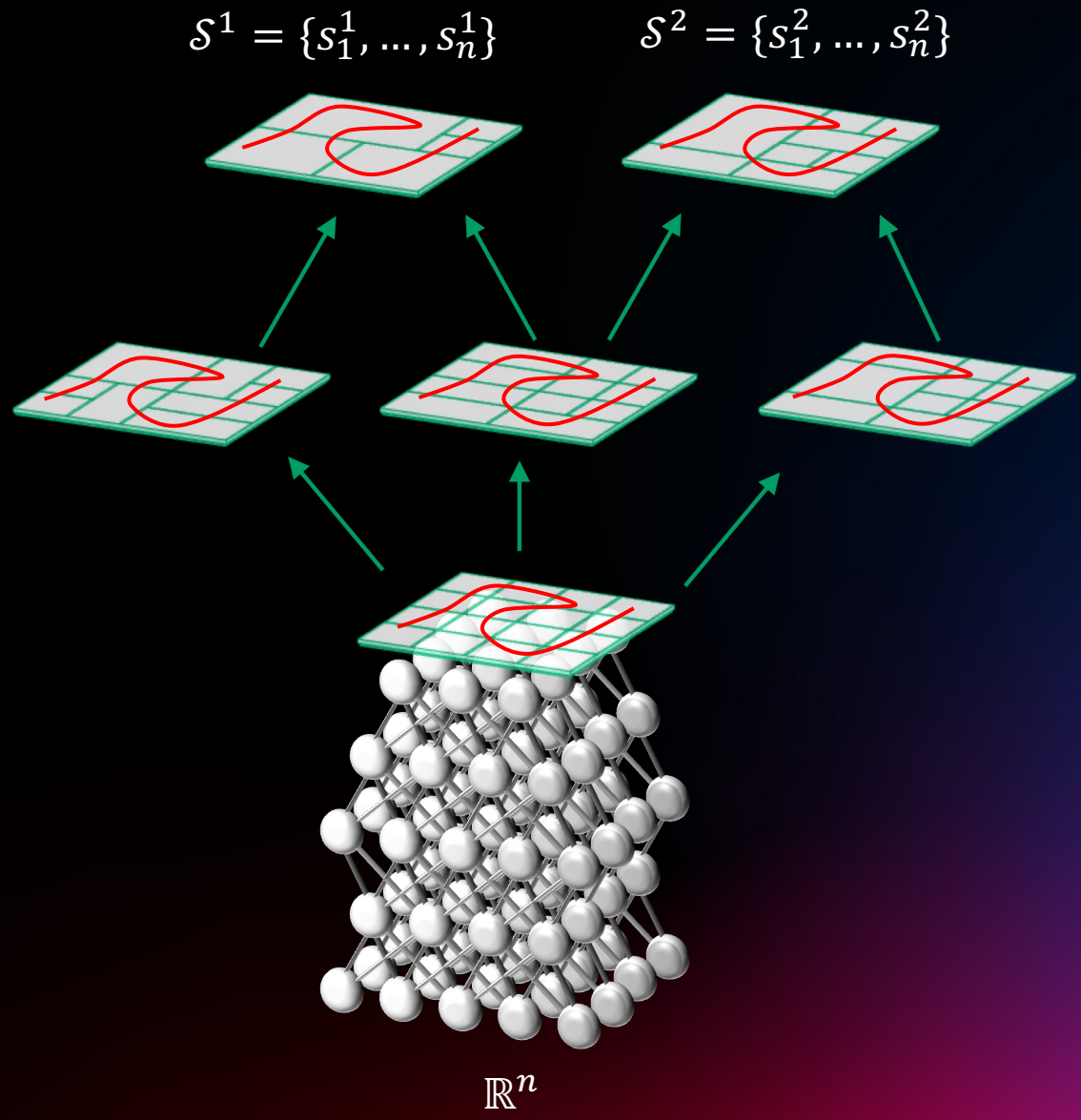
Challenges:

- Development of abstractions that are consistent
- Composability properties of symbolic abstractions:
  how do we engineer them? Is it an emergent property?
- Integrate interpretability requirements



Uncertainty and Ambiguity

- Neuro-symbolic representations are, in general, going to require a “quantization” of the concrete continuous space (e.g., state/actions) into symbols
- In general, this mapping is going to be lossy inducing both quantization error (uncertainty) as well as a non-uniqueness (ambiguity)

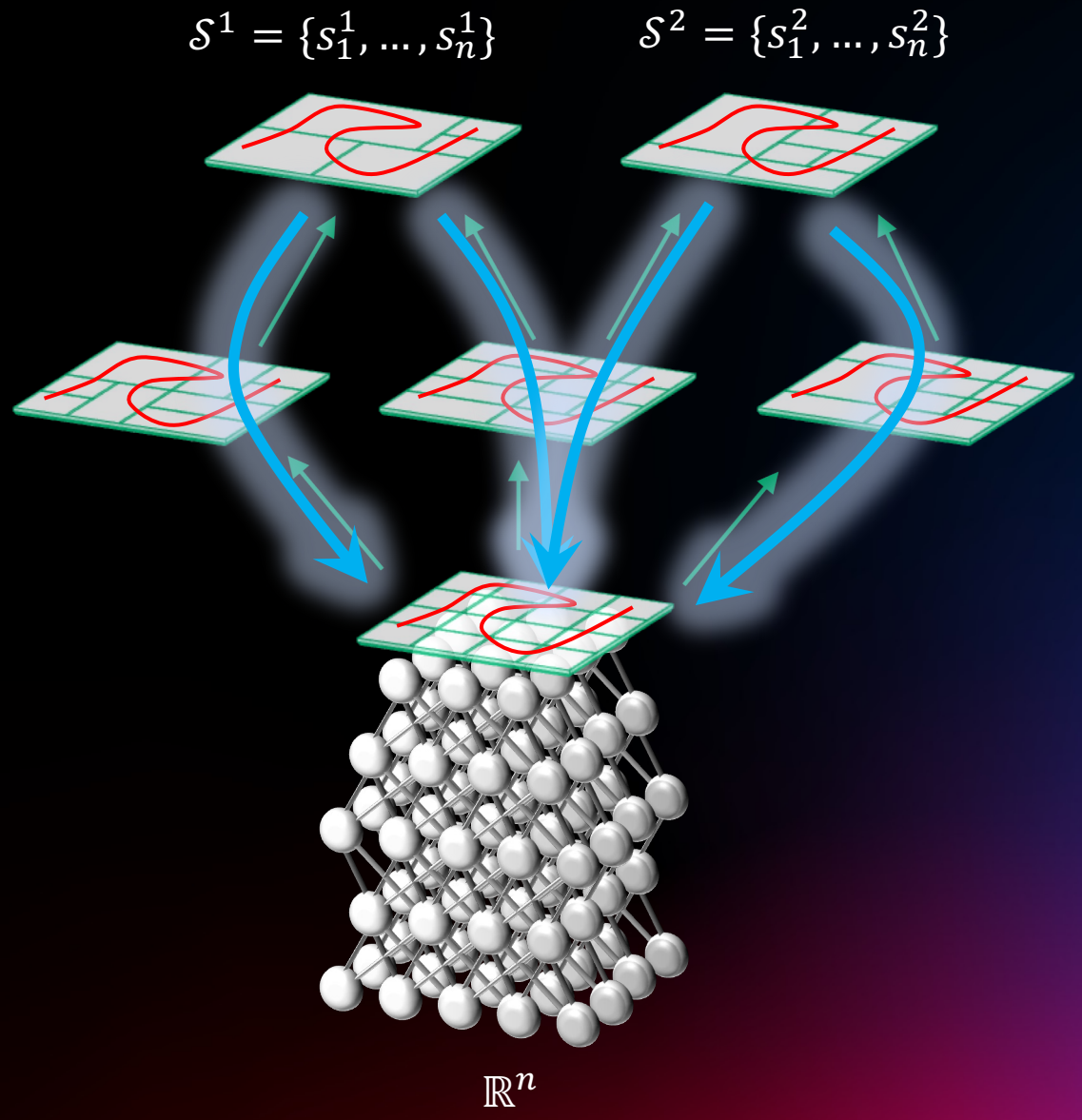


Uncertainty and Ambiguity

- Neuro-symbolic representations are, in general, going to require a “quantization” of the concrete continuous space (e.g., state/actions) into symbols
- In general, this mapping is going to be lossy inducing both quantization error (uncertainty) as well as a non-uniqueness (ambiguity)

Challenges:

- Formal way of incorporating quantization error and ambiguity in the symbolic abstractions (e.g., information theoretical approaches)
- The grounding of symbols is a well known problem but in an OODA loop we need to ensure consistency of symbols across the components (e.g., “Observe” and “Orient”)



Assurance of Complex Systems

Overarching Properties[1]:

- **Intent**: The system's *defined intended behavior* must be *correct and complete* with respect to the desired behavior
- **Correctness**: The *implementation must be correct* with respect to the intent under foreseeable operating conditions
- **Innocuity**: Unintended behavior must *not have unacceptable impact*
- **Operation**: The system must possess *mechanisms for addressing correctness* or *intent deficiencies* and for mitigating unacceptable impacts manifested during operation

https://nittanyai.psu.edu/events/lockheed-martin-wildfires-machine-learning-and-the-technology-of-the-future/

[1] C. M. Holloway, "Understanding the Overarching Properties". NASA, Langley Research C

Assurance of NS Systems

Challenges & Opportunities:

- **Intent:** NS representations enable to (more) formally describe the behavior of an autonomous system, at least at the level of each subcomponent of the OODA loop
- **Correctness:** Compositional NS abstractions provide opportunities to leverage contract based design methods[2]; however, uncertainty, ambiguity and symbolic consistency across OODA loop components need to be addressed
- **Innocuity**: Safety concerns, typically described as requirements in English, can be translated into (modal) logic and thus incorporated within NS architectures
- **Operation:** NS abstractions are built by incorporating logical formulas that can be monitored and reasoned ove during execution; however ambiguity, symbol grounding and uncertainty issues need to be addressed

https://nittanyai.psu.edu/events/lockheed-martin-wildfires-machine-learning-and-the-technology-of-the-future/

[2] A. Sangiovanni-Vincentelli, W. Damm, R. Passerone, "Taming Dr. Frankenstein: Contract-Based Design for Cyber-Physical Systems", European Journal of Control, 2012

What is needed to build trust-worthy and trusted AI

Example: Mission Planning through Preference Learning

Approach: Combine strengths of automated route planners with experience and intuition of pilots using interactive learning to produce high quality personalized route

Compared to a baseline semi-manual route planning, the Preference Learning

- Produced final routes more preferred by SMEs
- Resulted in higher favorable attitude ratings by SMEs
- Lowered subjective SME workload

Needs:

- Manage AI performance/explainability trade-off
- AI-based human attention management (when and how to engage humans)
- AI agents and humans learning from each other (build trust over time)
- Immersive virtual environments for joint human-AI training

Trust-worthy does not imply trusted

Directions for future research

- The OODA loop provides unique challenges for NS architectures but they also provide huge opportunities to design **trustworthy autonomous systems**

- **Composition, interpretability and reusability** of neuro-symbolic abstractions in the OODA loop can unlock new and more scalable design paradigms

- More broadly we need methods to account for:
  - Multiple spatio-temporal scales
  - Life-long learning at the edge
  - Human-AI trust building
  - Ethical AI
  - Multiple assurance frameworks (multiple suppliers)