Facilitating Rapid Teaming via Common Grounding

Prepared for Dr. Eric Davis

Susmit Jha
Technical Director
Neuro-symbolic Computing and Intelligence Research Group
Computer Science Laboratory
SRI

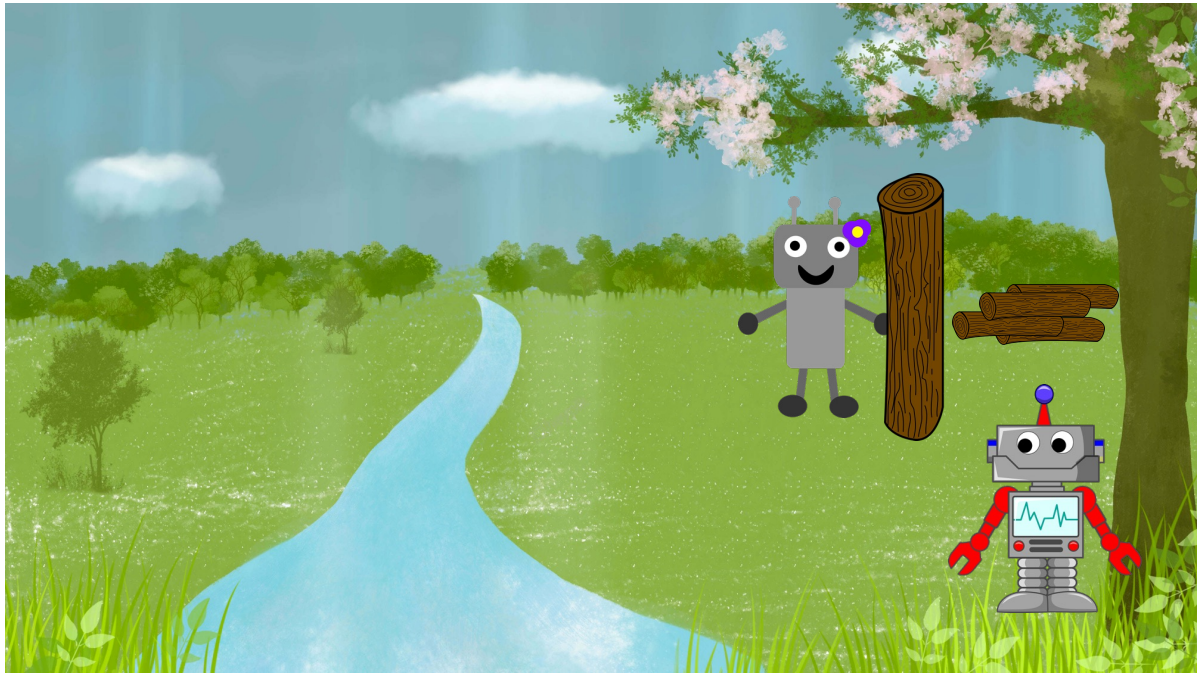
Facilitating Teaming using AI-guided Common Grounding

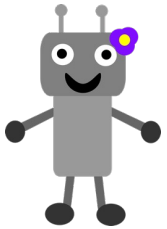
- **Understanding** This involves building a **causal world model**:
  - Taking into account the **consequences of one's own actions** on the
    - Environment
    - Other's beliefs and preferences
  - **Predicting other's behaviour with counterfactual** analysis on
    - Implications of other's beliefs and preferences on their behavior
    - Evolution of intentions and preferences

- **Communication**. The ability to explicitly and credibly share information with others relevant to understanding behaviour, preferences/desire and intentions.
  - **Commitment**. The ability to make credible promises (e.g. respect for preferences, execution of an intent) when needed for cooperation.
  - **Norms and institutions**. Social infrastructure — such as **shared beliefs or rules** — that reinforces understanding and communication.

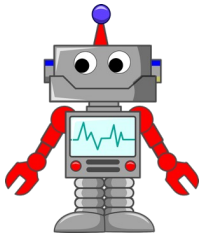
susmit.jha@sri.com

Interpretable Learning for Shared Intentionality





Alice



Bob

Humans can undertake novel, collective behavior, or teamwork
Capability to communicate goals, plans and ideas to create shared intentionality

Consider two autonomous agents Alice and Bob with cognition capability.

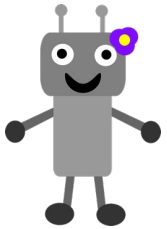
Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

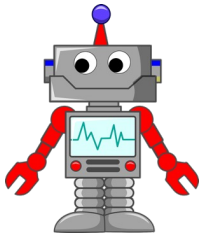
Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

susmit.jha@sri.com

Interpretable Learning for Shared Intentionality



Alice



Bob

- Beyond Propositional to Temporal Behavior
- Noisy trajectories

Consider two autonomous agents Alice and Bob with cognition capability.

Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

susmit.jha@sri.com

Interpretable Learning for Shared Intentionality

Consider two autonomous agents Alice and Bob with cognition capability.

Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

susmit.jha@sri.com

Interpretable Learning for Shared Intentionality

Consider two autonomous agents Alice and Bob with cognition capability.

Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

susmit.jha@sri.com

Interpretable Learning for Shared Intentionality

Consider two autonomous agents Alice and Bob with cognition capability.

Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

susmit.jha@sri.com

Interpretable Learning for Shared Intentionality

Consider two autonomous agents Alice and Bob with cognition capability.

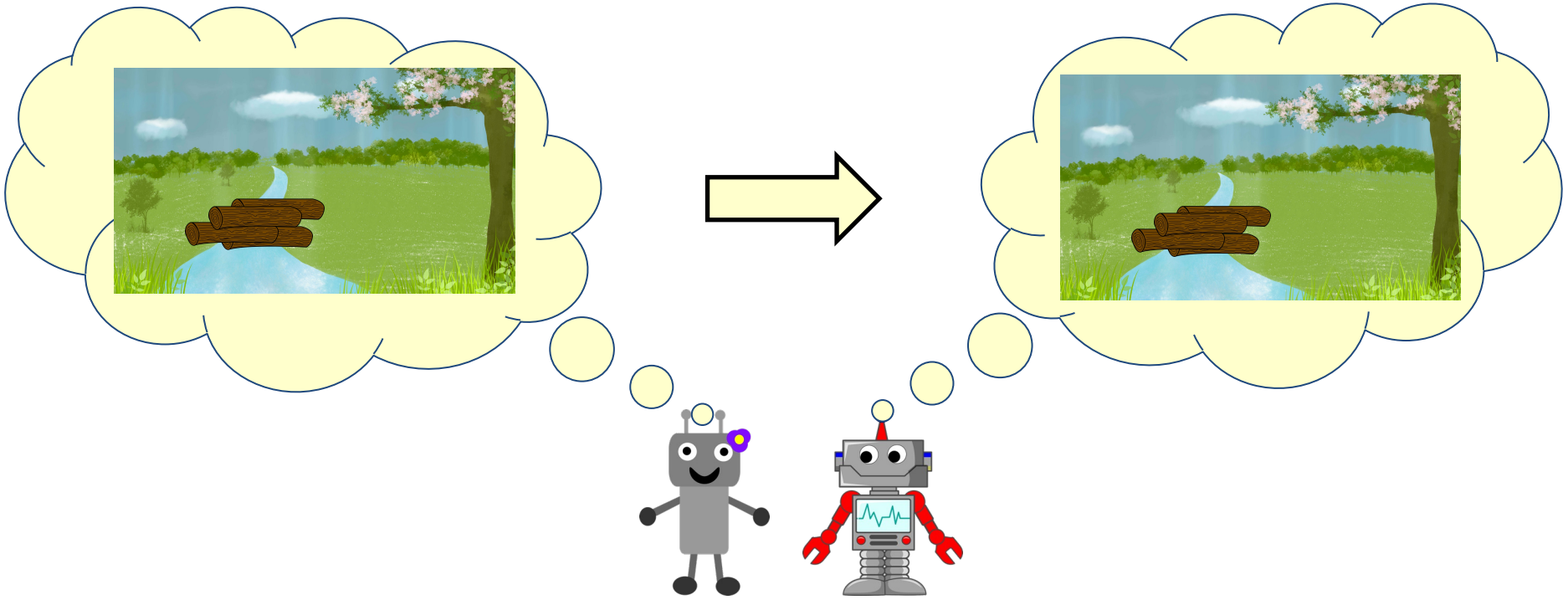
Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

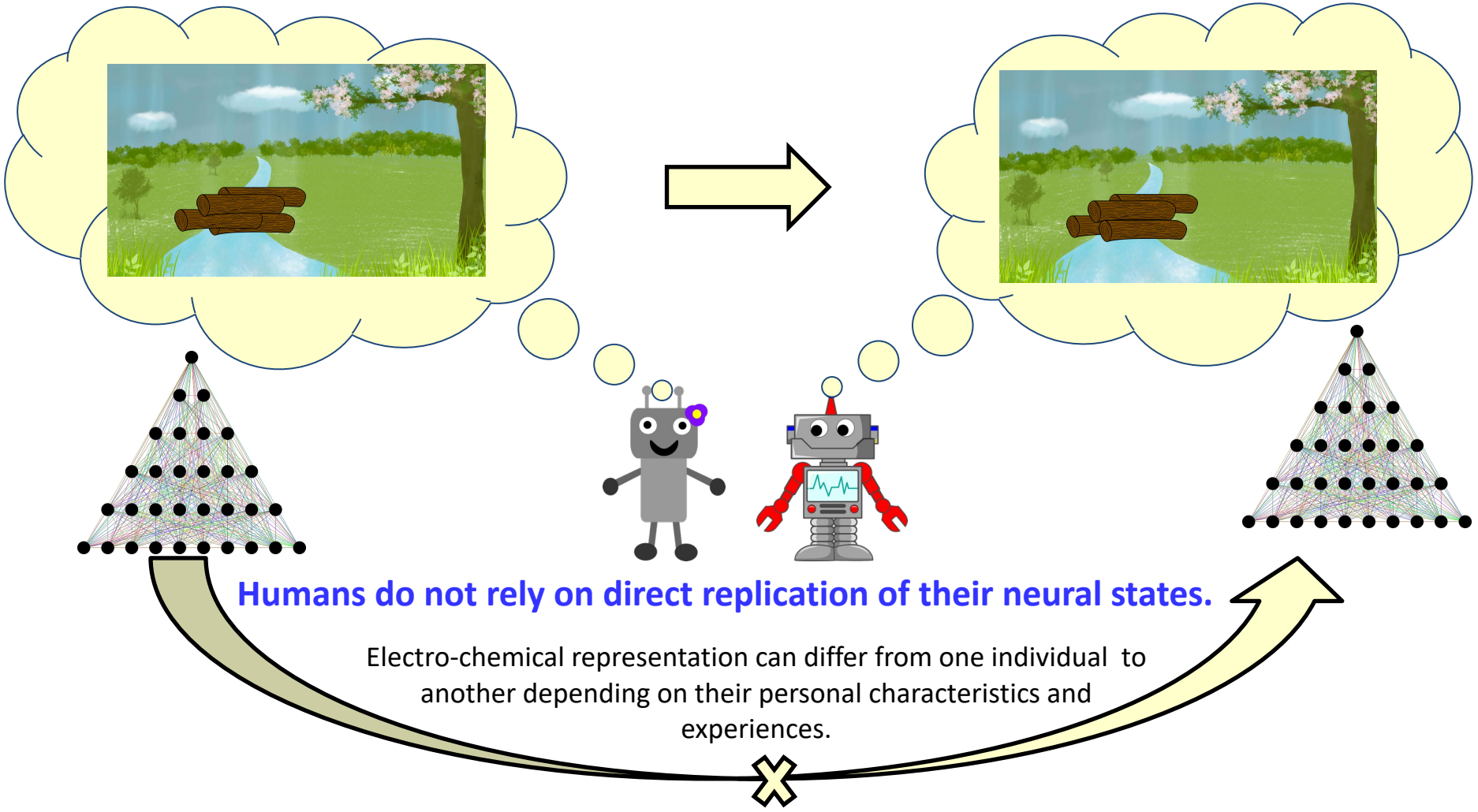
susmit.jha@sri.com

Interpretable Learning for Shared Intentionality



Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

Shared Intentionality: Mental Cloning?

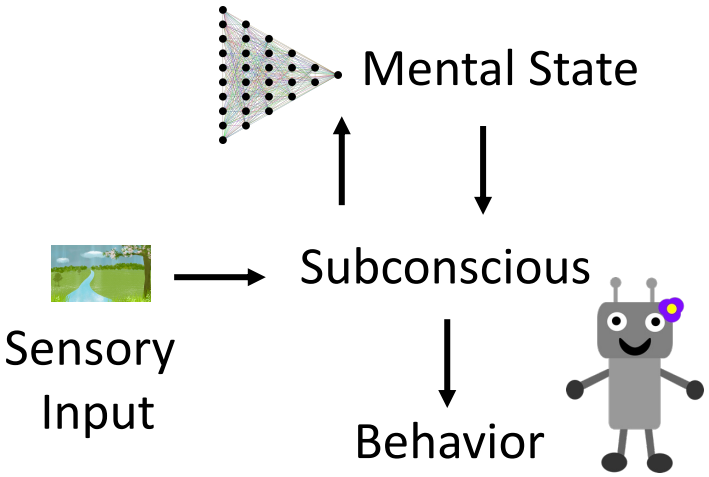


Humans do not rely on direct replication of their neural states.

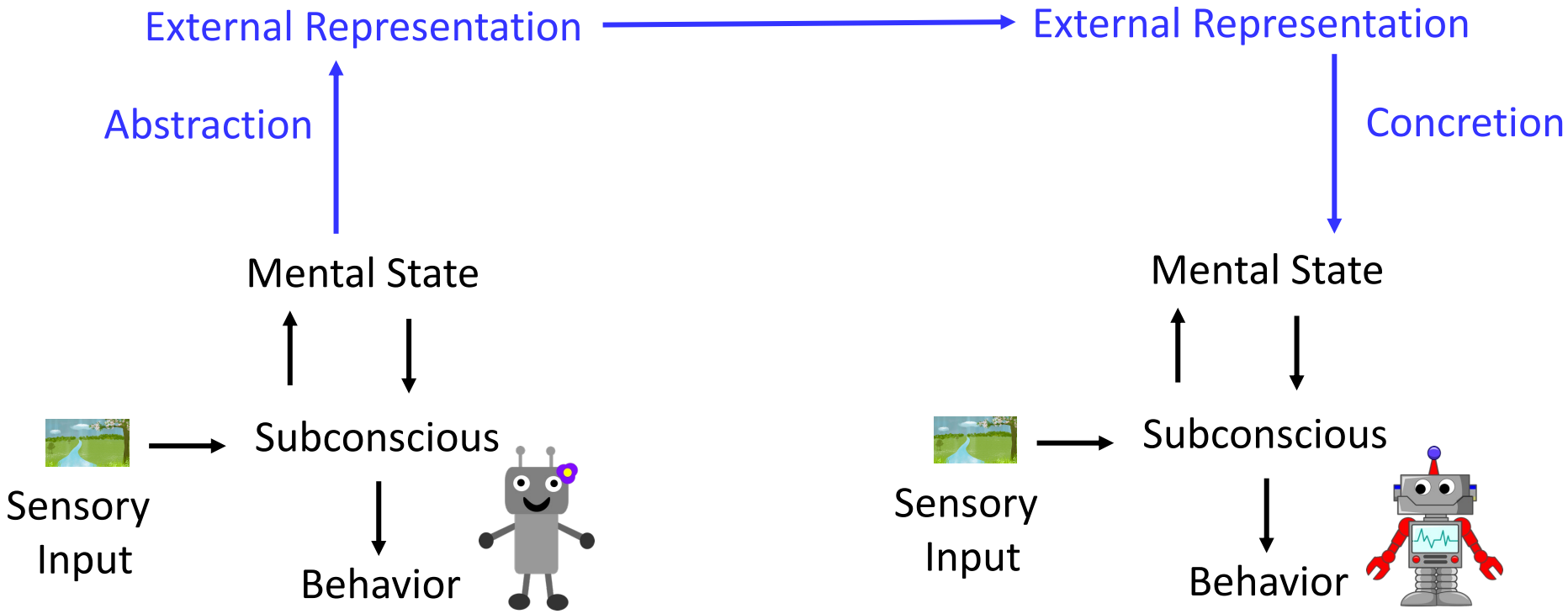
Electro-chemical representation can differ from one individual to another depending on their personal characteristics and experiences.

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice.

Cognitive Architecture: Subconscious



Cognitive Architecture: Communication

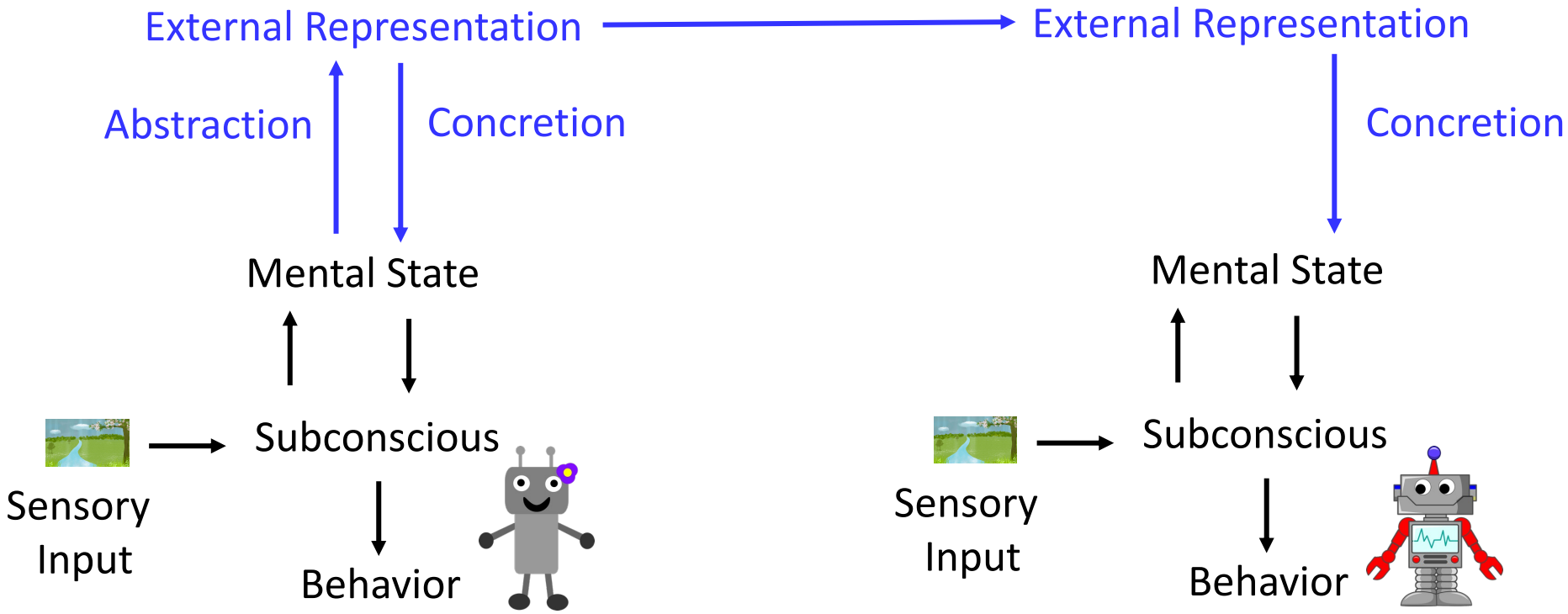


Alice needs to construct an explanation – a succinct external communication – whose concretion by Bob will reproduce parts of Alice's subconscious mental state.

Evans, J. (2003). "In two minds: dual-process accounts of reasoning". *Trends in Cognitive Sciences*

Cognitive Architecture: Theory of Mind

Gweon, H., Saxe, R. (2013). Developmental cognitive neuroscience of Theory of Mind. *Neural Circuit Development and Function in the Brain: Comprehensive Developmental Neuroscience.*

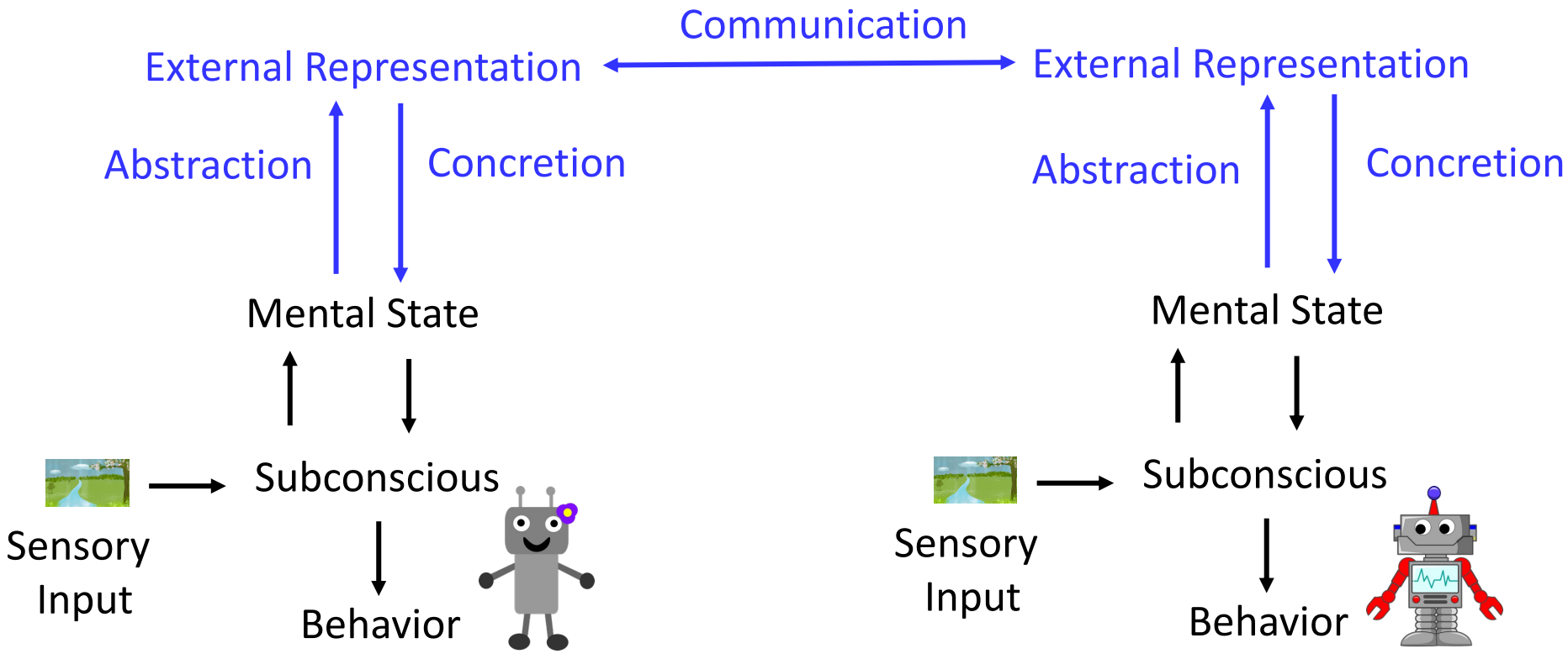


Alice needs to construct an explanation – a succinct external communication – whose concretion by Bob will reproduce parts of Alice's subconscious mental state.

- Create suitable mental model of Bob's brain and his concretion operator
- Use that to construct suitable abstraction/explanations to be communicated

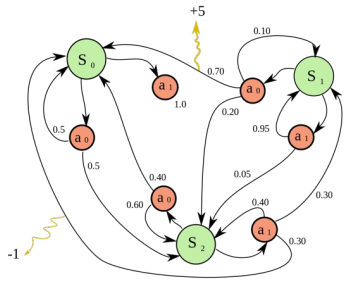
Cognitive Architecture: Dual Process Theory

Evans, J. (2003). "In two minds: dual-process accounts of reasoning". *Trends in Cognitive Sciences*



To explain an idea: Alice constructs her abstraction and explanation so that when concretized using her model of Bob's concretization function she gets back to her own subconscious state. This is done by predictive processing : iterate to reduce prediction error.

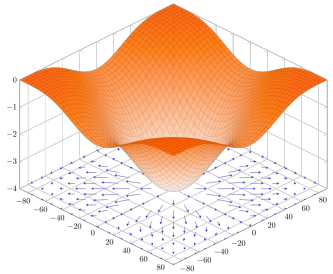
Communicating Using Demonstrations: Neuro-symbolic Inverse Reinforcement Learning (IRL) [Neurips'18]



Environment Markov Decision Process

Expert Demonstrations

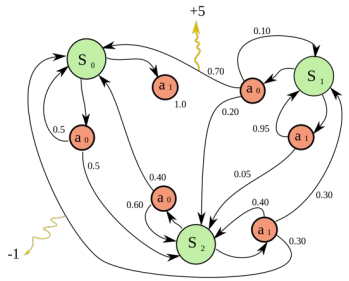
Inverse Reinforcement Learning



Numerical Reward Function

- Non-composable (not intentions but low-level numerical weights)
- Non-interpretable
- Difficult to transfer and adapt to new contexts

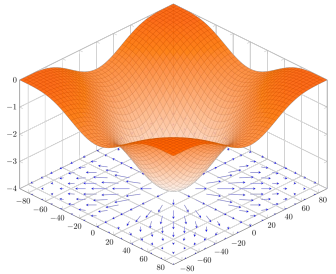
Communicating Using Demonstrations: Non-Markovian IRL [Neurips'18]



Environment Markov Decision Process

Noisy Expert Demonstrations

Inverse Reinforcement Learning



Numerical Reward Function

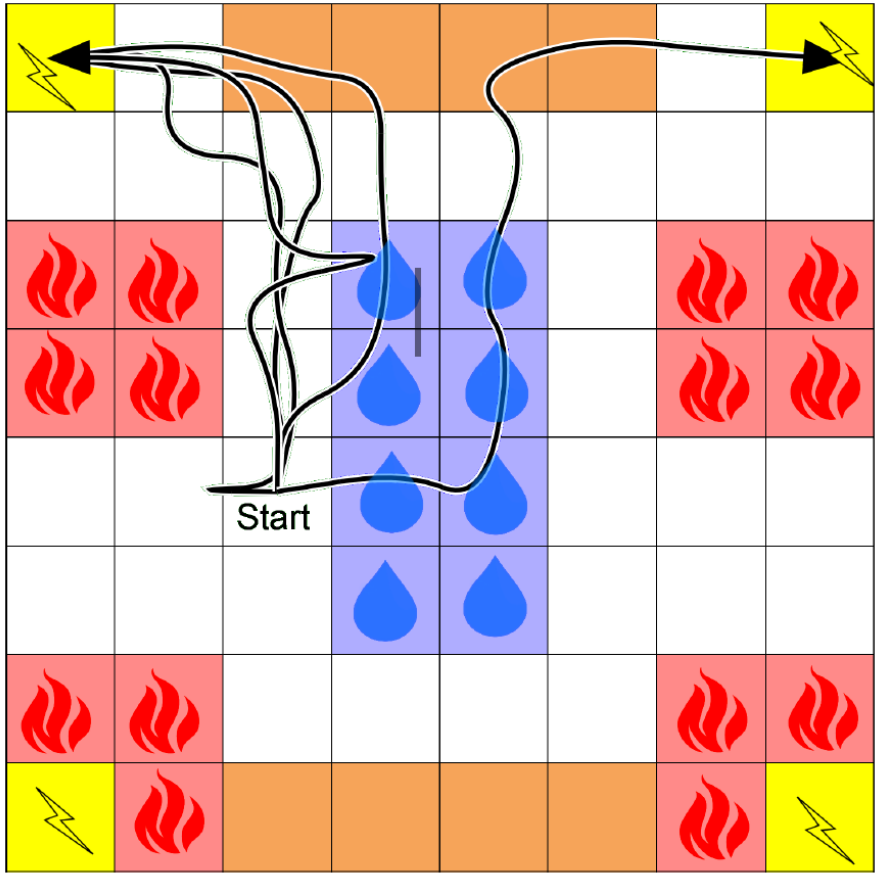
- Demonstrations and rewards are often non-Markovian due to mental state of the actor not directly modeled by environment MDP.

Communicating Using Demonstrations: Grid World Example

1. Avoid fire (red).

2. Eventually Recharge (yellow).

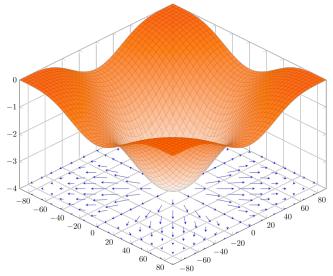
3. If you touch the water (blue) then dry off (brown) before recharging (yellow).



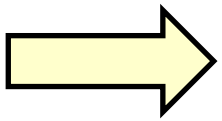
Explicit reduction to non-Markovean representation suffers from the curse of history.

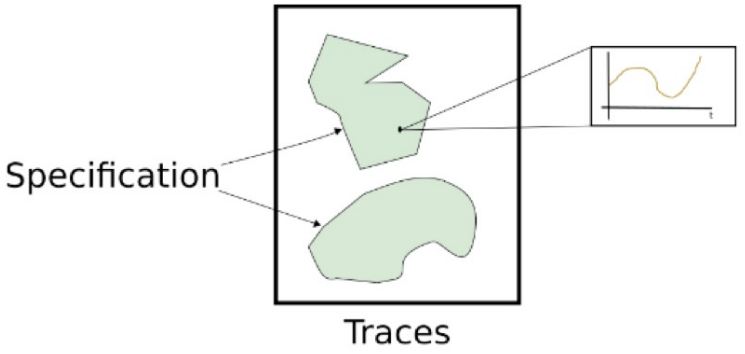
a. (4 colors)^(10 time steps) = 2^20 traces ≈ 1048576
b. #specifications = 2^(2^20) ≈ 10^315652

Communicating Using Demonstrations: Temporal logic specifications



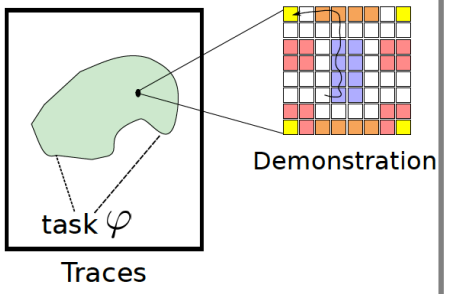
Numerical Reward Function

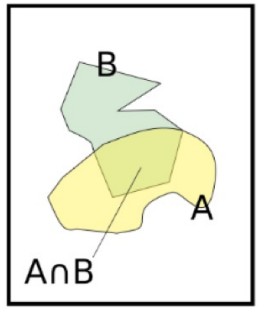


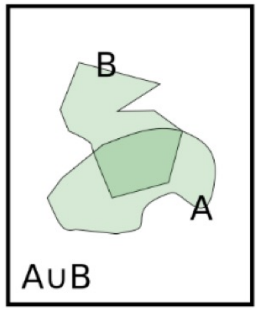


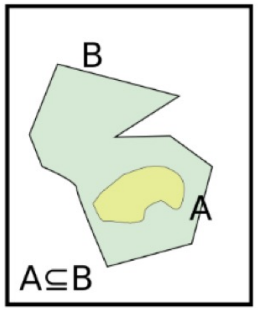
Trace Properties
as Task Specifications

$$r_\varphi(\xi) = \begin{cases} 1 & \text{if } \xi \in \varphi \\ 0 & \text{otherwise} \end{cases}$$



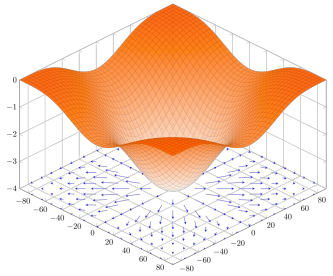




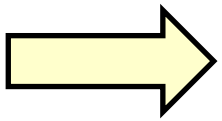


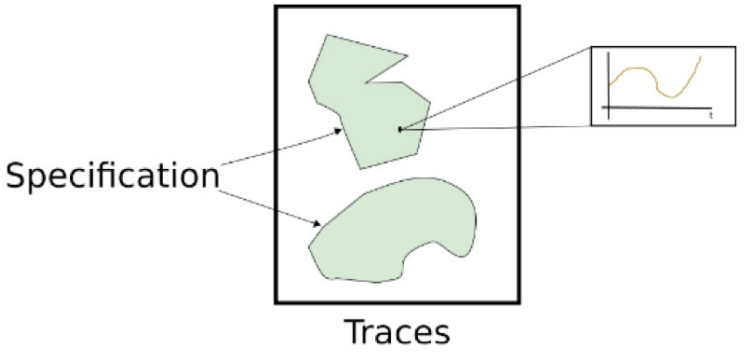
- Composable
- Resilient to changes in task context
- Interpretable
- Can leverage formal methods tools

Communicating Using Demonstrations: Temporal logic specifications



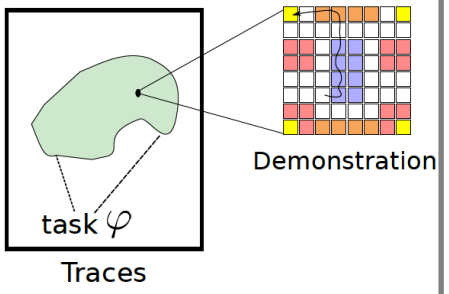
Numerical Reward Function

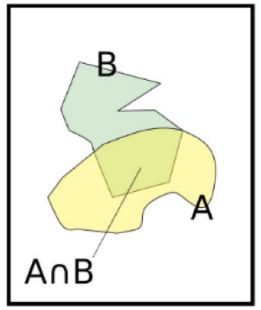


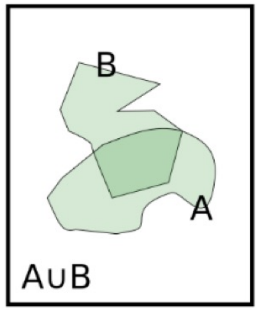


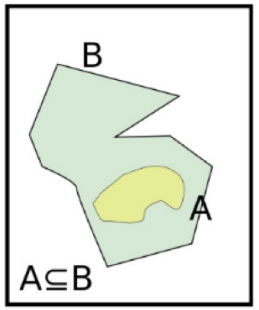
Trace Properties as Task Specifications

$$r_\varphi(\xi) = \begin{cases} 1 & \text{if } \xi \in \varphi \\ 0 & \text{otherwise} \end{cases}$$





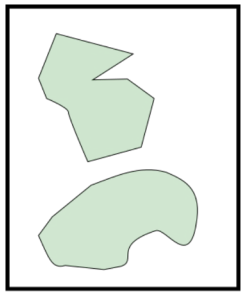
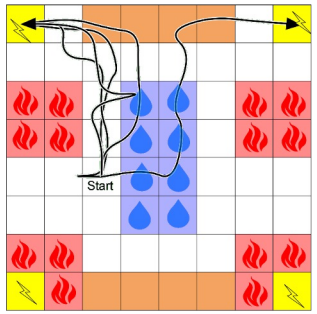




- Pnueli, Amir. "The temporal logic of programs." IEEE, 1977.
- Donzé, Alexandre, and Oded Maler. "Robust satisfaction of temporal logic over real-valued signals." *FORMATS*, 2010.
- Jha, Susmit, Vasumathi Raman, Dorsa Sadigh, and Sanjit A. Seshia. "Safe autonomy under perception uncertainty using chance-constrained temporal logic." *Journal of Automated Reasoning* 60, 2018.

Communicating Using Demonstrations: Specification Inference Problem

Like most inverse problems, this problem is underspecified.

What is Pr( | 

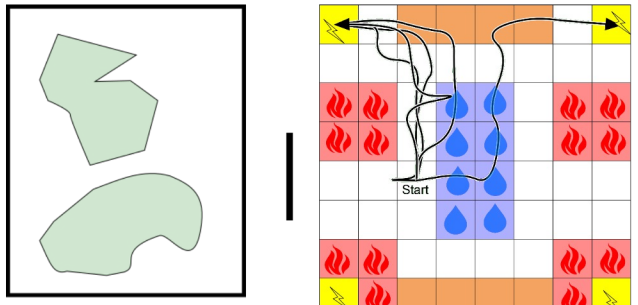
) ?

Specification

- Intent satisfaction is Boolean. Either Alice/Bob did the task or didn't.
- Assuming Alice is at least better at performing the task than a random action policy.
- Applying the principle of maximum entropy select the the distribution.
  - Inspired by Maximum Entropy Principle (also used in Inverse Reinforcement Learning)

Communicating Using Demonstrations: KL Divergence

$$\Pr(\ \text{Specification}\ \mid\ \text{Demonstrations}\ ) \propto e^{D_{KL}\left(\mathcal{B}(\overline{\varphi}) \parallel \mathcal{B}(\widehat{\varphi})\right)}$$



Bernoulli Distribution

Satisfaction probability for Alice given dynamics

Satisfaction probability given uniformly random actions

Marcell Vazquez-Chanlatte, Susmit Jha , Ashish Tiwari, Mark K. Ho and Sanjit A. Seshia. Learning Task Specifications from Demonstrations. NeurIPS, 2018

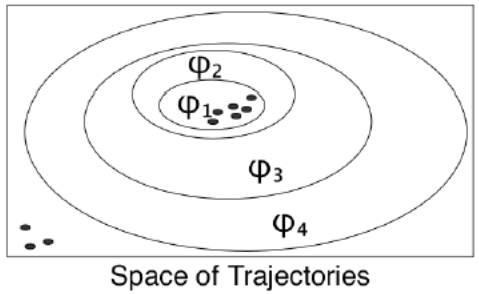
Communicating Using Demonstrations: Computing posterior

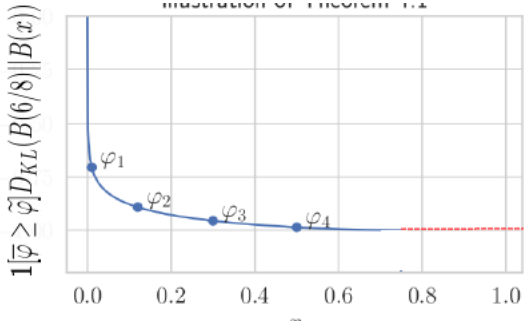
Maximum a Posteriori

$$\max_{\varphi} \; D_{KL}\left(\mathcal{B}(\overline{\varphi}) \parallel \mathcal{B}(\widehat{\varphi})\right)$$

Algorithm Sketch

If one fixes the measured sat probability, the KL-divergence term in the model is convex in the random satisfaction rate. This enables an efficient lattice based search for the most probable specification.



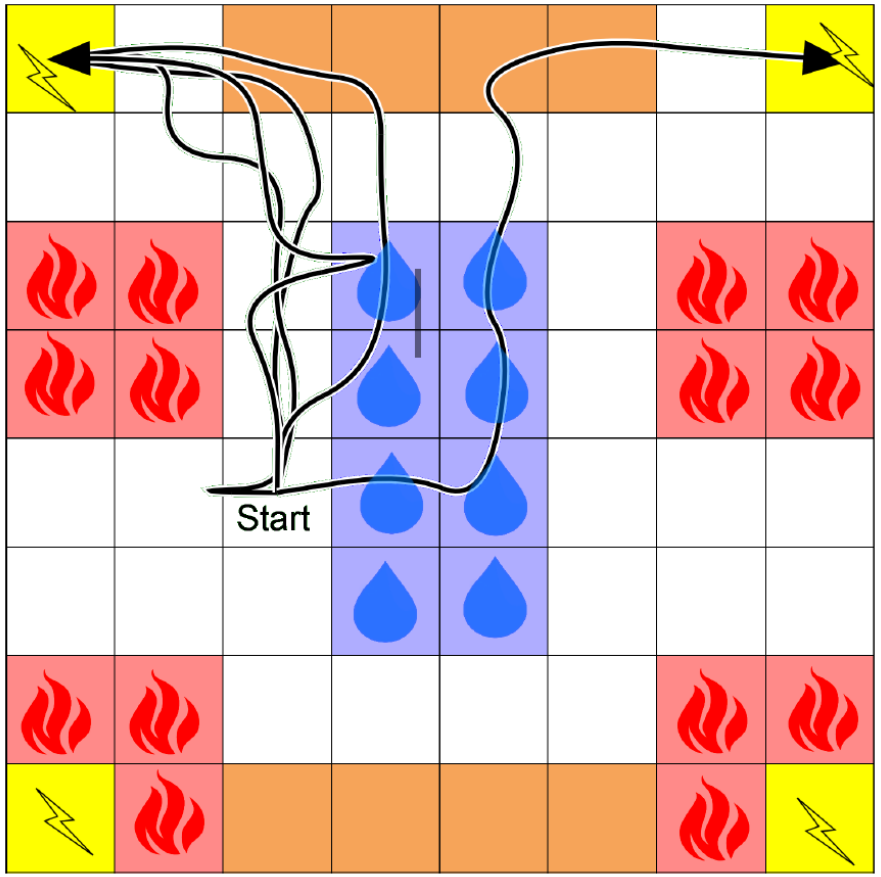


Communicating Using Demonstrations: More involved example

1. Avoid fire (red).

2. Eventually Recharge (yellow).

3. If you touch the water (blue) then dry off (brown) before recharging (yellow).

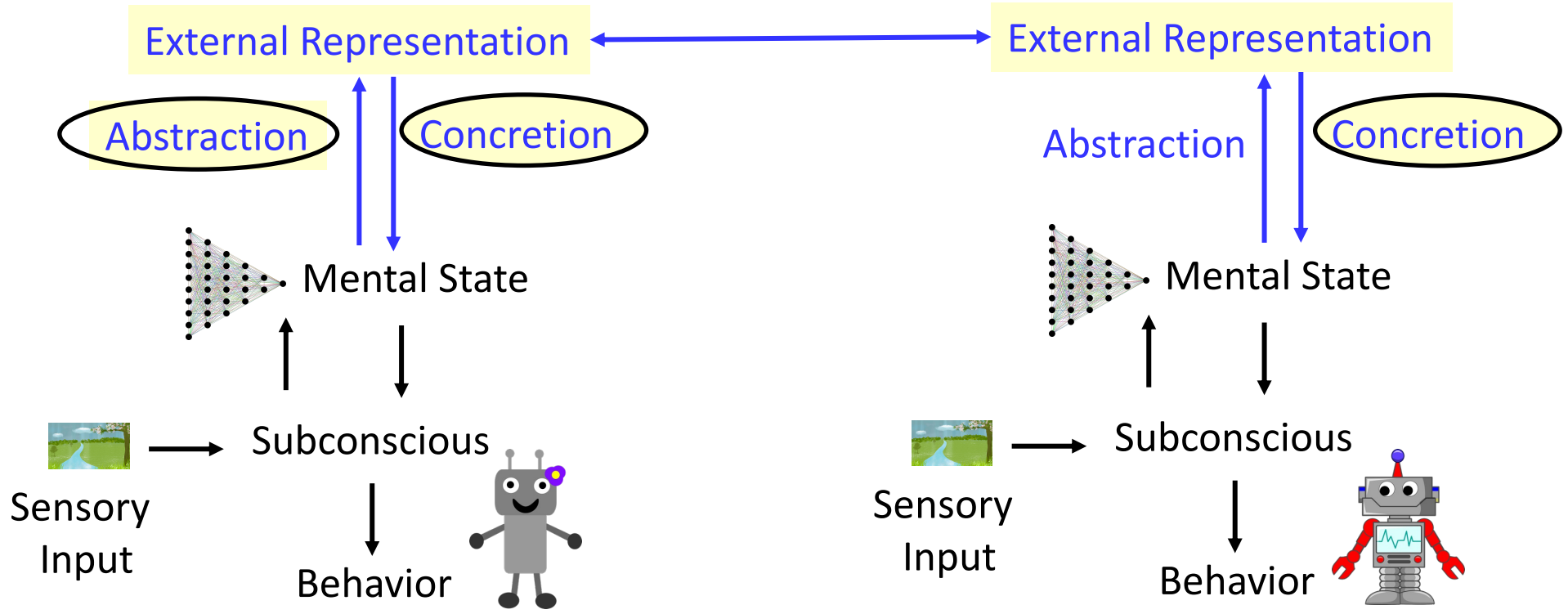


Temporal Logic Specification

H: Historically
O: Once
S: Since

$$\left(H\neg red \wedge O\ yellow\right) \wedge H\left(\left(yellow \wedge O\ blue\right) \Rightarrow \left(\neg blue\ S\ brown\right)\right)$$

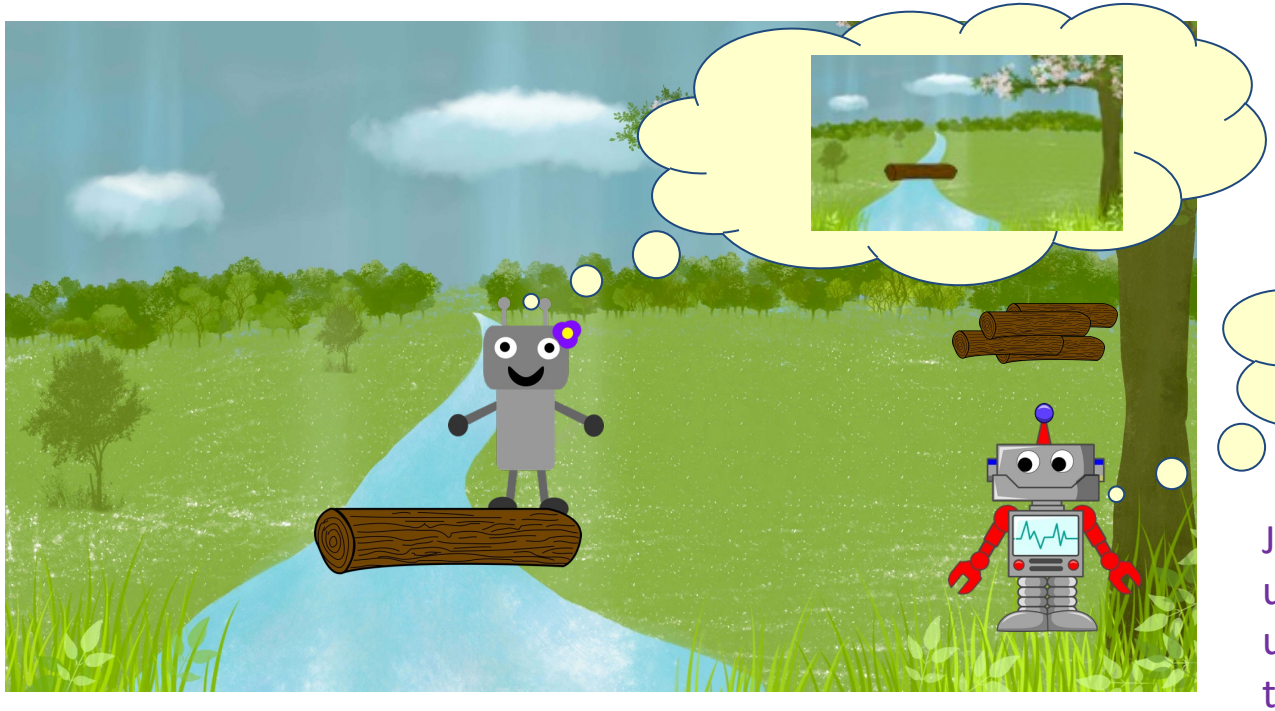
Cognitive Architecture: Communication



Implementing communication via Logical Specification Inverse Reinforcement Learning

A Candidate Mechanism to Computationally Implement Shared Intentionality

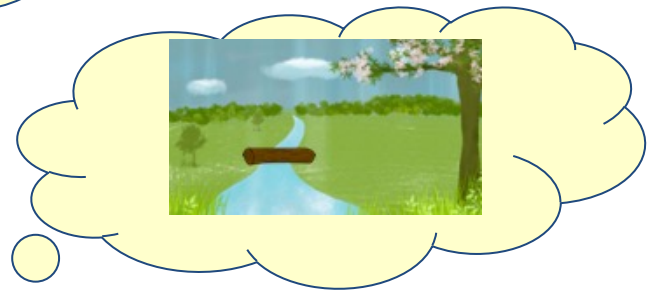
Inferring and Conveying Intentionality: Beyond Numerical Rewards to Logical Intentions. Susmit Jha and John Rushby. AAAI Spring Symposium, Towards Conscious AI Systems, 2019



Find Specification as Maximum a Posteriori

$$\max_{\varphi} D_{KL}\left(\mathcal{B}(\overline{\varphi}) \parallel \mathcal{B}(\widehat{\varphi})\right)$$

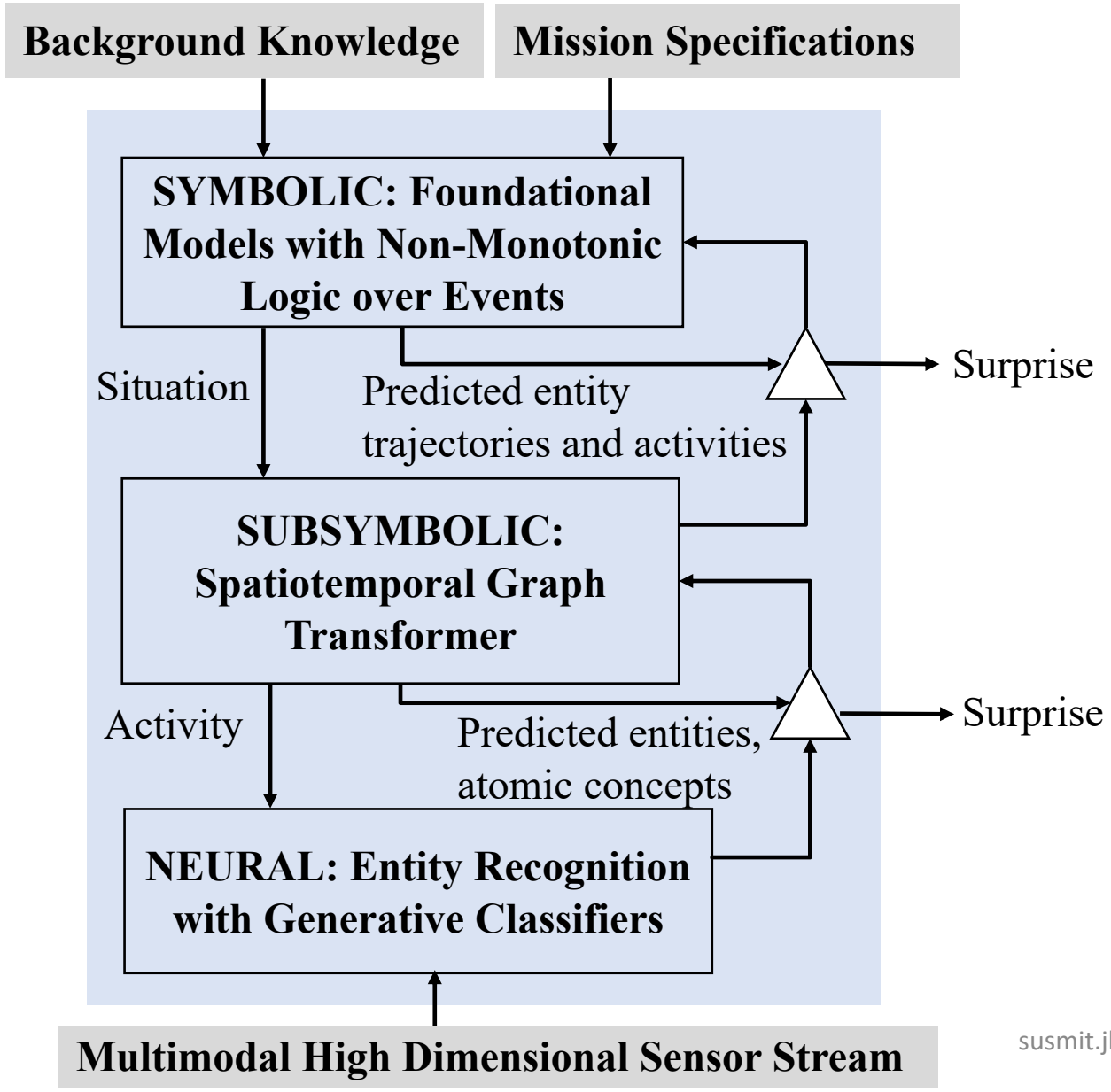
Marcell Chanlatte, Susmit Jha , Ashish Tiwari, Mark K. Ho and Sanjit A. Seshia. Learning Task Specifications from Demonstrations. NeurIPS, 2018



Jha, Susmit et al. "Safe autonomy under perception uncertainty using chance-constrained temporal logic." *Journal of Automated Reasoning* 60, 2018

susmit.jha@sri.com

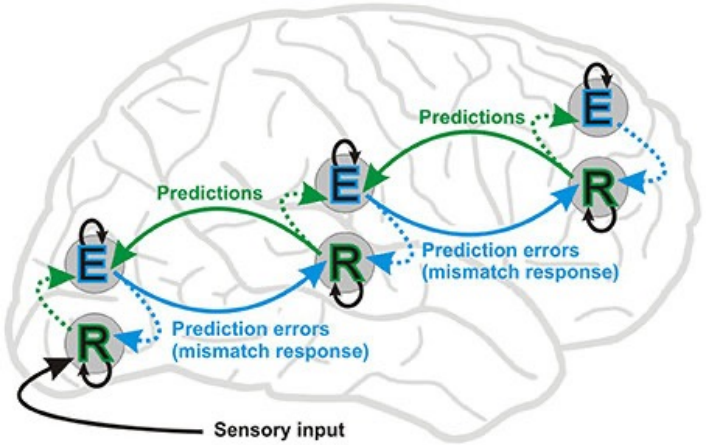
AI Agent Interlocutor $(AI)^2$ - An instance of BDI (Belief-Desire-Intentions) Architecture



Predictive coding (also known as predictive processing) is **a theory of mind in which the mind is constantly generating and updating a mental model of the environment**. The model is used to generate predictions of sensory input that are compared to actual sensory input.

Rao and Ballard'99, Friston and Kiebel'09

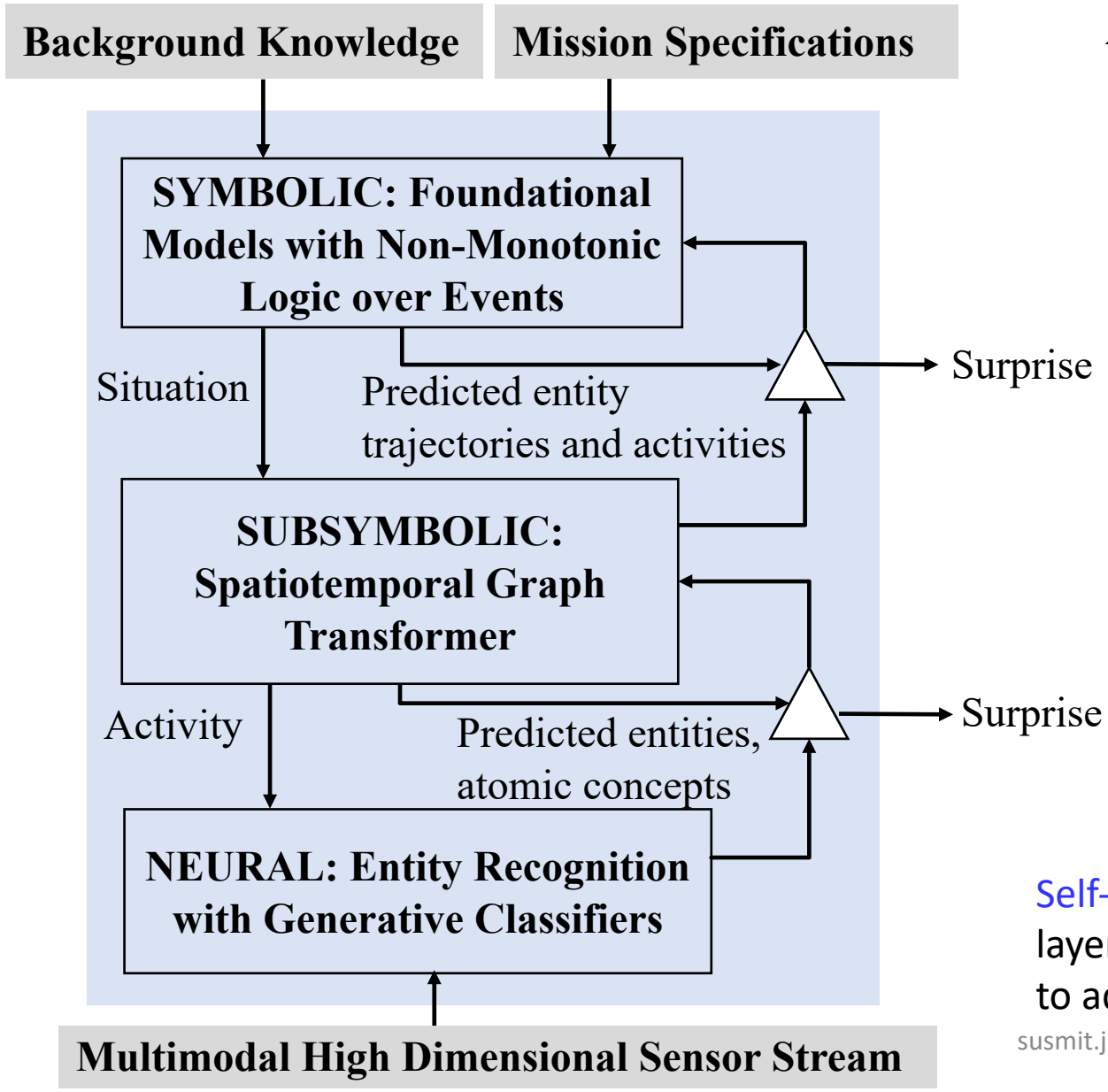
Stefanics et. al.'14



susmit.jha@sri.com

DARPA Assured Autonomy, DARPA ANSR

AI Agent Interlocutor $(AI)^2$ - An instance of BDI (Belief-Desire-Intentions) Architecture



Incorporate common sense and general knowledge

Constructing abstract concepts

Predicting using *larger* contexts

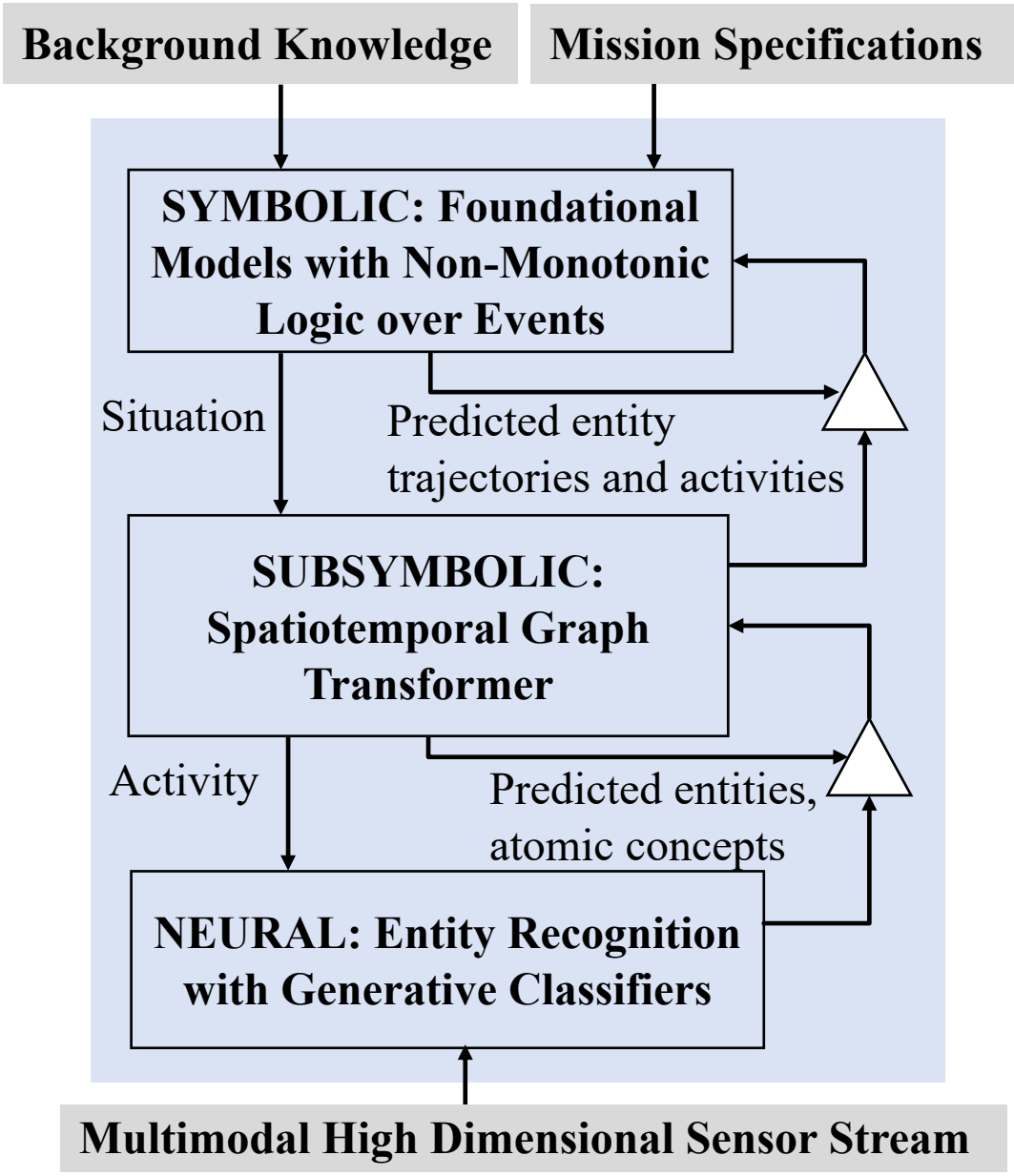
Self-stabilizing loops across layers makes TrinityAI robust to adversarial perturbations.

- Attribution-Based Confidence (ABC) Metric For Deep Neural Networks. Jha et. al. NeurIPS, 2019
- Model-Centered Assurance for Autonomous Systems. Jha et. al. 39th International Conference on Computer Safety, Reliability and Security (SafeComp), 2020
- On Detection of Out of Distribution Inputs in Deep Neural Networks. Jha et al. IEEE CMI 2021
- iDECODe: In-distribution Equivariance for Conformal Out-of-distribution Detection. Kaur et. al. AAAI 2022
- Detecting out-of-context objects using graph contextual reasoning network. Acharya et. al. IJCAI, 2022.
- Cunningham et al, Principal Manifold Flows, ICML'22
- Kaur et. Al. CODiT: Conformal Out-of-Distribution Detection in Time-Series Data for Cyber-Physical Systems. ICCPS'23
- Gupta et al. Class Prototypes based Contrastive Learning for Classifying Multi-Label and Fine-Grained Educational Videos. CVPR'23

susmit.jha@sri.com

DARPA Assured Autonomy, DARPA ANSR

AI Agent Interlocutor $(AI)^2$ - An instance of BDI (Belief-Desire-Intentions) Architecture



We learn the abstract dynamics over the events as a logic program over the atomic events. Background knowledge is incorporated as default rules in the logic program.

Two key characteristics:

- Negation as failure: Instead of predicate logic in which a predicate can be either true or false, open world reasoning needs to take into account ”do not know”
- Nested Exceptions: Rules and learned dynamics might not remain consistent as situation evolves and our logical framework allows exceptions (with arbitrary depth of nesting).

Trinity so far has largely focused on what can be called “**first-order understanding**” and does not model the recursion of how one agent’s beliefs can be about the beliefs of others, which themselves may be about the beliefs of the first, and so on.

susmit.jha@sri.com

DARPA Assured Autonomy, DARPA ANSR

AI Agent Interlocutor $(AI)^2$ - An instance of BDI (Belief-Desire-Intentions) Architecture

Identification of distinct atomic events

- Identifying distinct events or states from the provided observations

Realizing the temporal order

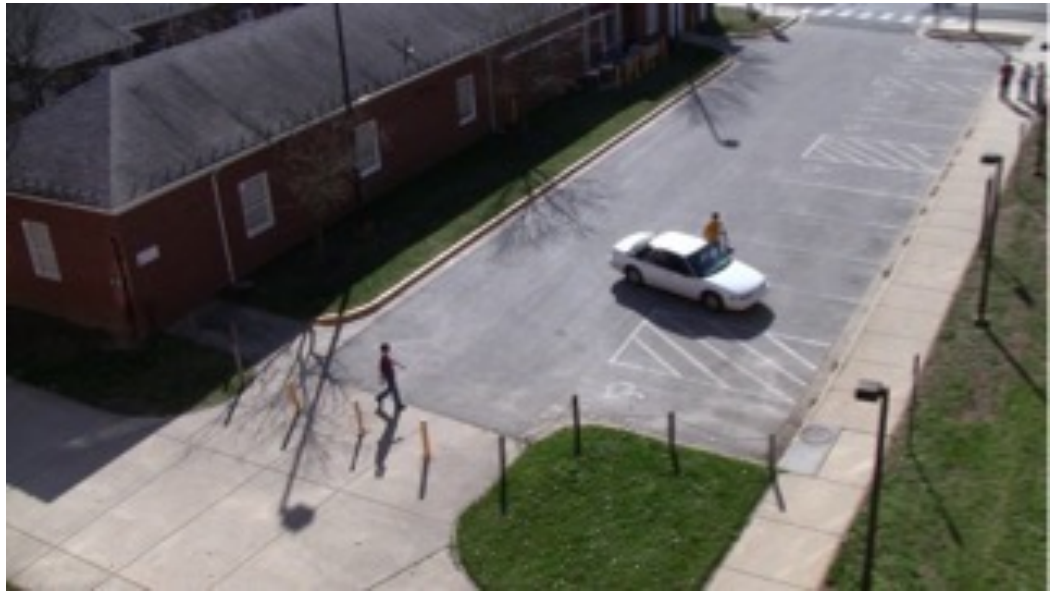
- Generate a temporal order between the observations

Inferring causal relationships

- Infer the cause-and-effect relationships considering unobserved events

Iterative refinement

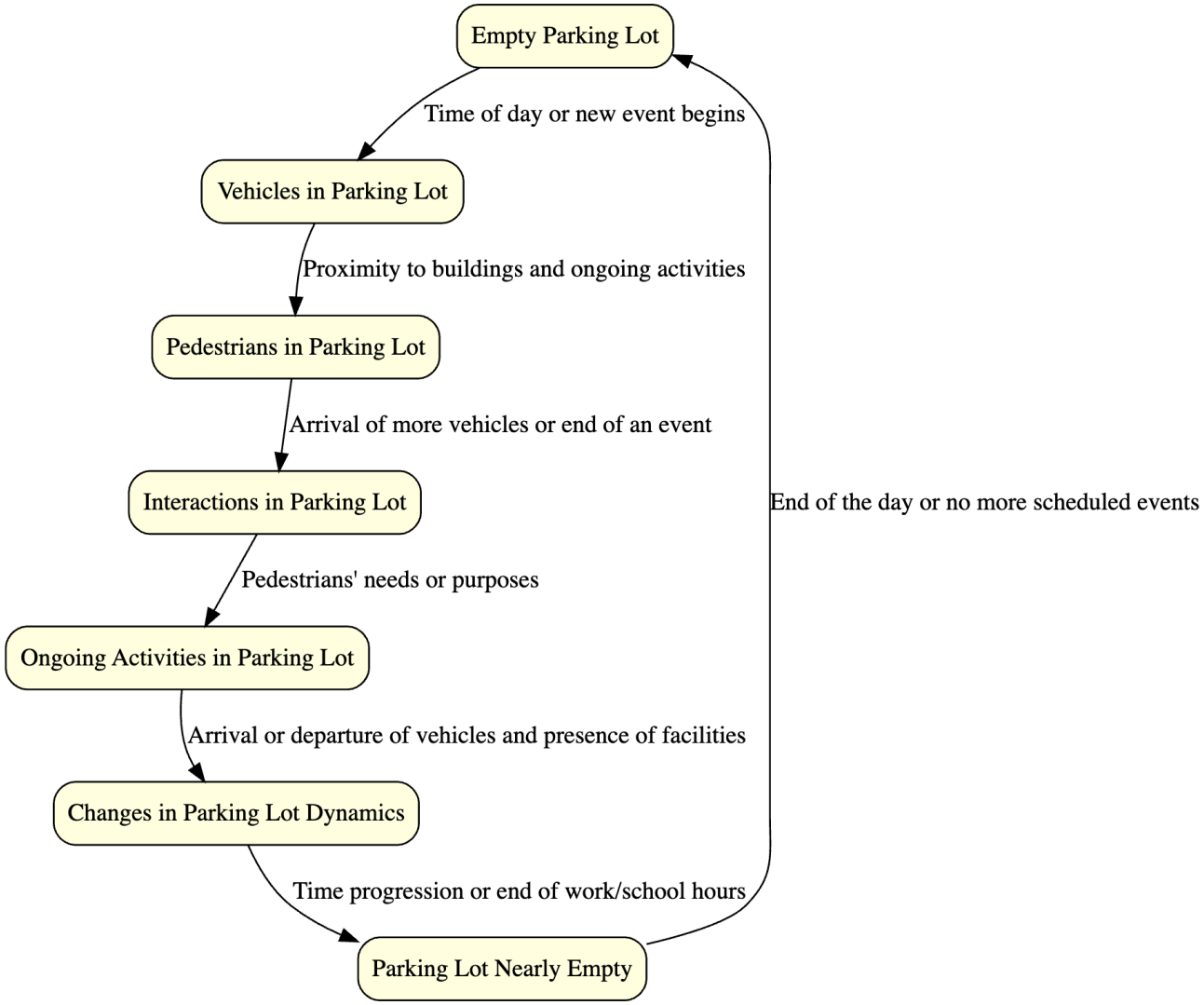
- Revisiting the observations to ensure that the derived causal relationships are consistent with the observations.



susmit.jha@sri.com

DARPA Assured Autonomy, DARPA ANSR

AI Agent Interlocutor $(AI)^2$ - An instance of BDI (Belief-Desire-Intentions) Architecture



Causal State Transition:

1.State: Empty Parking Lot
- **Transition:** Time of day or new event begins.
- **Resulting State:** Vehicles start entering the parking lot.

2.State: Vehicles in Parking Lot
- **Transition:** Proximity to buildings (e.g., school or brick building) and ongoing activities.
- **Resulting State:** Pedestrians start to populate the area.

3.State: Pedestrians in Parking Lot
- **Transition:** Arrival of more vehicles or end of an event.
- **Resulting State:** Increase in foot traffic and interactions.

4.State: Interactions in Parking Lot
- **Transition:** Pedestrians' needs or purposes, like waiting for a taxi or conversing with others.
- **Resulting State:** Specific activities such as a taxi driver waiting for passengers or individuals conversing.

5.State: Ongoing Activities in Parking Lot
- **Transition:** Arrival or departure of vehicles and the presence of facilities like a pedestrian sign.
- **Resulting State:** Changes in parking lot dynamics, such as searching for parking spots or increased foot traffic.

6.State: Changes in Parking Lot Dynamics
- **Transition:** Time progression or end of work/school hours.
- **Resulting State:** Parking lot starts to empty out.

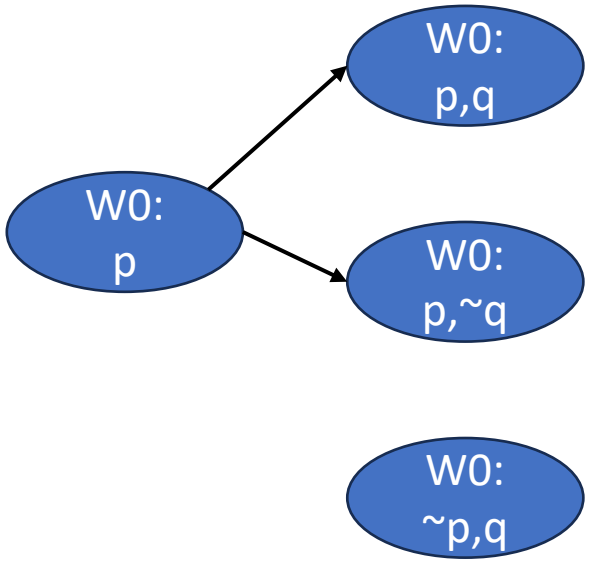
7.State: Parking Lot Nearly Empty
- **Transition:** End of the day or no more scheduled events.
- **Resulting State:** Parking lot completely empty.

susmit.jha@sri.com

DARPA Assured Autonomy, DARPA ANSR

Modal Logic and Possible-world Semantics

- Modal Logic approach combines possible worlds semantics for knowledge and belief (Kripke)
- Possible worlds semantics for actions, planning and prediction (Moore, Fagin, Halpern, Morgenstern, Davis)
- Allows representation of nested beliefs and reasoning over beliefs (E.g. Tom understands that Sam's action X is interpreted by Jerry to indicate intent Y)
- Difference between knowledge and belief: real world is not accessible to agent in case of belief.

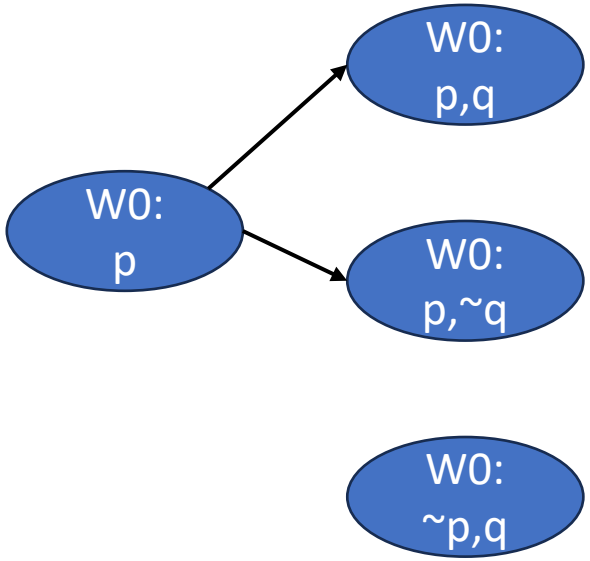


Teaming requires aggregating belief graphs and making them consistent.

susmit.jha@sri.com

Modal Logic and Possible-world Semantics

- Need to combine causal world model with reasoning over beliefs.
- Rieter's improved version of the situation calculus is a classical temporal logic that can be used here.
- Actions in this framework will serve two purpose
  - Link "before" situations with "after" situations
  - Define epistemic accessibility relation between agents and situations (possible worlds)
- Enables reasoning over the world evolution as well as each agent's knowledge and belief over the world.
- Correctly being able to predict other agent's is an implicit learning signal (apart from explicit feedback)



Teaming requires aggregating belief graphs and making them consistent.

susmit.jha@sri.com

Conclusion

- Causal World Model (DARPA Assured Autonomy, DARPA CAML)
- Neuro-symbolic Inverse Reinforcement Learning (DARPA Assured Autonomy; also applied this to DARPA CAML)
  - NeurIPS'18, AAAI Symposium on Consciousness'19
- Modal Logics for Reasoning over Beliefs (DARPA ALIAS)
- Semi-supervised emergence of common language

susmit.jha@sri.com