Event Memory for Explainable Autonomy

David Menager and Dongkyu Choi

Who we are

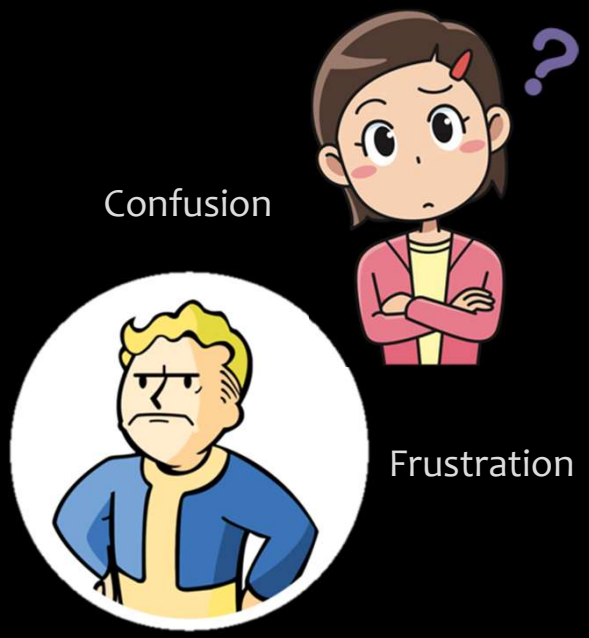
David Menager
- PhD student at the University of Kansas
- Studies event memory in cognitive systems
- E-mail: dhmenager@ku.edu

Dongkyu Choi
- Senior scientist at Agency for Science, Technology and Research, Singapore
- Studies cognitive architectures and intelligent agents for physical domains
- E-mail: choid@scei.a-star.edu.sg

Agenda

- Motivation and background
- Hybrid theory of event memory
- Implications for autonomous agents
- Future work
- Conclusions

Motivation



Problem: No persistence over time

Solution: Enable agents to remember history of interaction using an event memory

Motivation

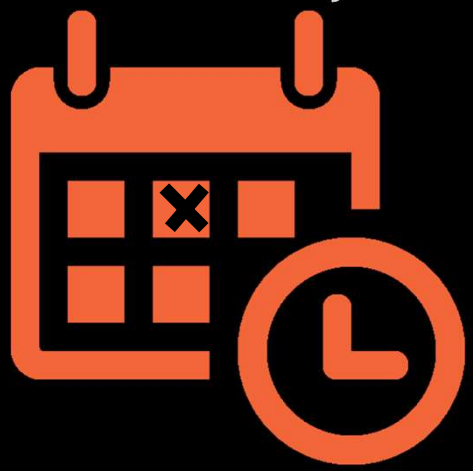
Problem: No capacity to explain decisions

Solution: Enable agents to justify their behaviour using an event memory

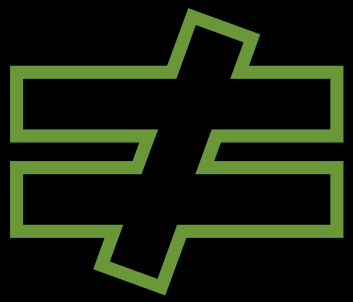
Background

Conflicting psychological theories of event memory

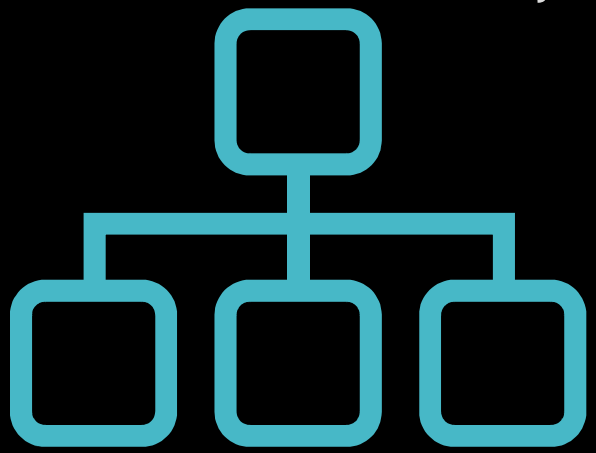
Causal Theory



Memory stores instances of events.



Simulationist Theory



Memory stores schemas of events.

Hybrid Theory of Event Memory

- The elements of event memory are organized in a hierarchical manner such that higher-level elements in the hierarchy contain probabilistic summaries of the lower-level ones;
- The memory for events is a long-term memory that stores episodes and event schema;
- Episodes are propositional representations of specific events whose contents are descriptions of the agent's perceptions, beliefs, goals, and intentions;
- Event schema are first-order propositional templates that summarize similar episodes or event schema in a probabilistic manner; and
- Remembering involves performing inference over the probability distributions contained in memory elements.

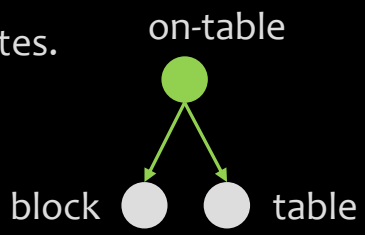
Percepts, Beliefs, and Intentions

- Agents perceive objects in the world as predicates with a type, a name, and attribute–value pairs.

  (block block1 x 2 y 0 width 2 height 2)
  (table table1 x 0 y 0 width 10)

- Agents infer beliefs about the world as relational predicates.

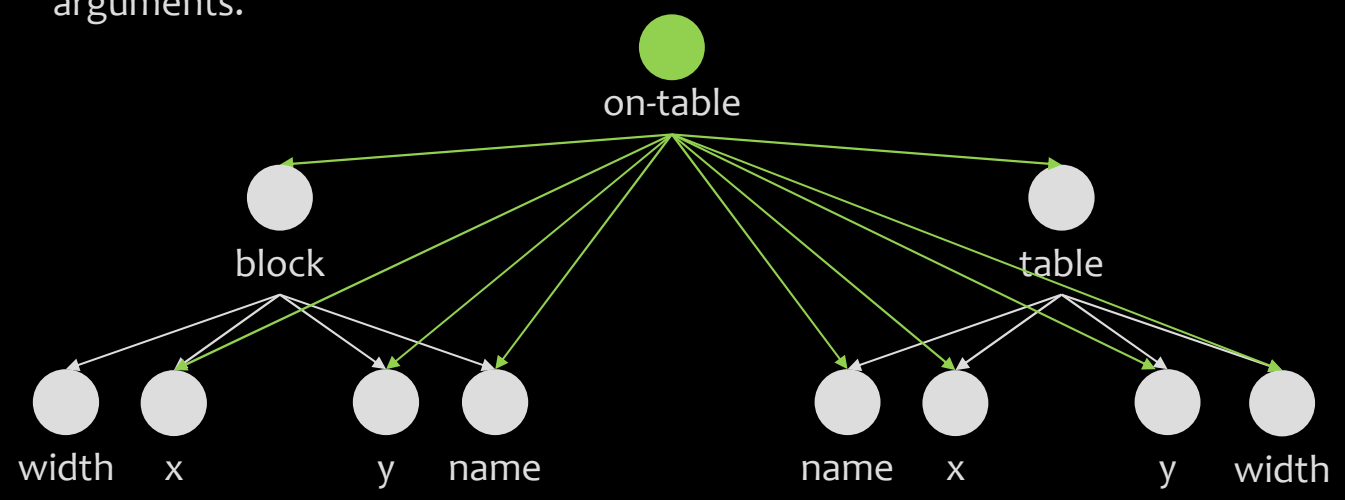
  (on-table block1 table1)



- Agents store their own intentions as state elements.

Episodes

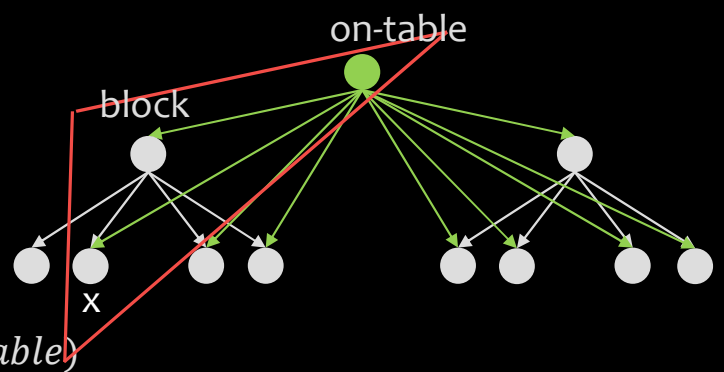
- Episodes are ordered sequences of states.
- A state defines a conditional independence relation over predicates and their arguments.



Event Schemas

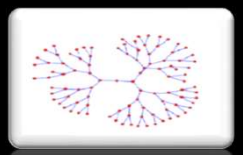
- Event schemas are probabilistic aggregates of episodes and other schemas.
- Event schemas add probability distributions to nodes in the belief state.

$$p(\boldsymbol{x}_{\{1:N\}}) = \prod_{i=1}^{n} p(x_i | Parents(x_i))$$

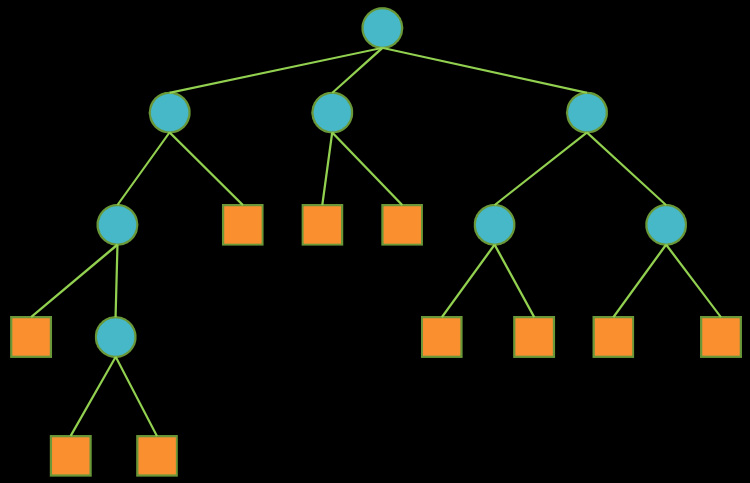


$$\mu_1, \mu_2$$
$$\sigma_1^2, \sigma_2^2$$
$$p(OnTable)$$
$$p(Block|onTable)$$
$$onTable \sim Bernoulli(pOnTable)$$
$$block \sim Bernoulli(pBlock|onTable)$$
$$\omega = [block * onTable, \quad block * (1 - onTable)]$$
$$z|\omega \sim Cat(\omega)$$
$$x|z \sim N(\mu_z, \sigma_z^2)$$

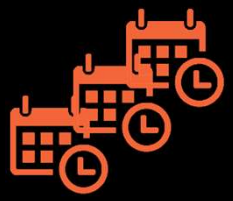
Event Memory Organization



- Our theory unifies aspects of the causal and the simulationist theories by employing hierarchy.
- Episodes are the leaf-level elements and are instances of actual events.
- There are layers of event schemata on top of the episodes at the leaf nodes.
- These schema have with characteristics of semantic memory elements that no longer maintain a causal link to a specific event.



Event Memory Process: Insertion

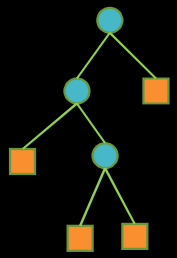


Given: episodic instances.

Form a partition that assigns observed instances into classes.

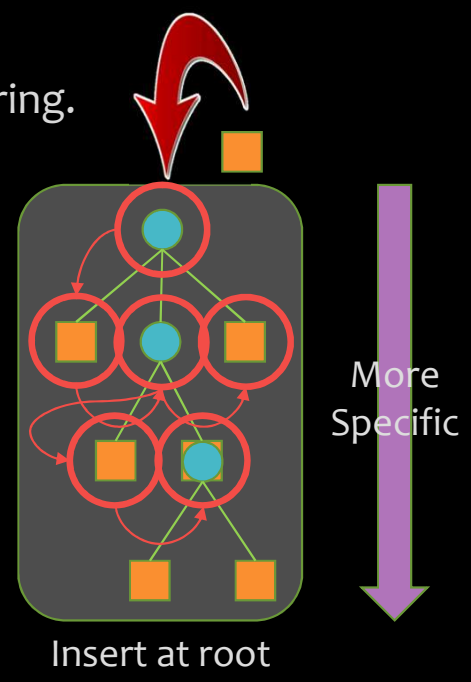
Build a probabilistic schema for each class.

Induce a hierarchical relation amongst the schemas and instances.



Event Memory Process: Insertion

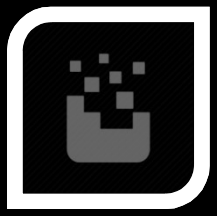
- Schemas and instances are stored in a hierarchy where each node is indexed by IS-A links from its parents.
- System simultaneously performs classification and clustering.
- Recursively checks instance with parent's children.
- Cost function is first-order propositional unification.



Event Memory Process: Remembering

Given: Retrieval Cue.

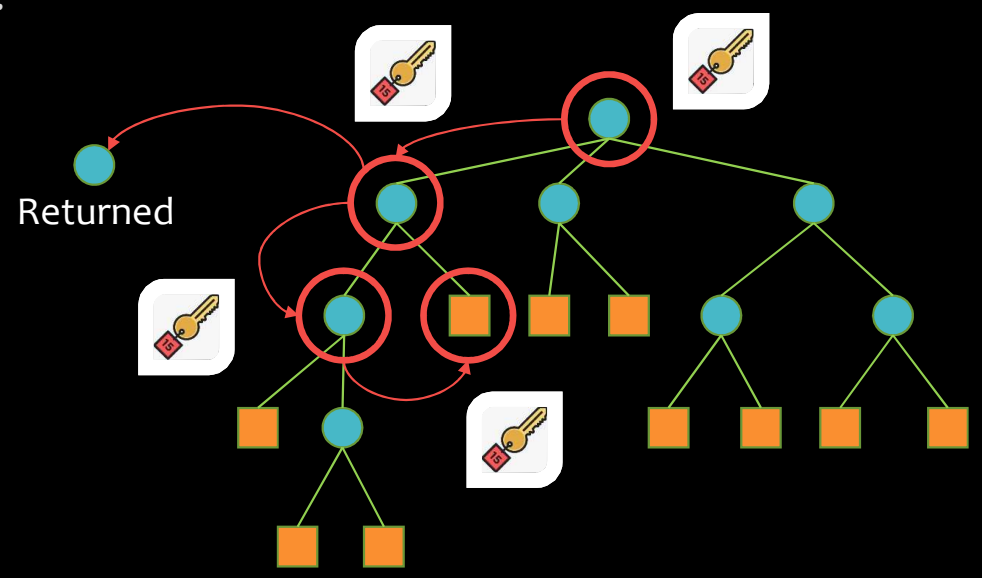
Find an episode or schema that best matches the cue.



Perform probabilistic inference on schemas to reconstruct the state.

Event Memory Process: Remembering

- The systems sorts the retrieval cue as if it was being inserted, but without updating probabilities in the schemas.
- At the end of the sort, the system returns best matching memory element and does not insert the cue.



Event Memory Process: Remembering

- The remembering process probabilistically infers the truth value of unobserved beliefs using a schema.
- Given a retrieval cue $\boldsymbol{x}_{cue}$, and unobserved variables $\boldsymbol{x}_h$, inference computes:

$$p(x_h | x_{cue}, \boldsymbol{\theta}) = \frac{p(x_h, x_{cue} | \boldsymbol{\theta})}{p(x_{cue} | \boldsymbol{\theta})}.$$

- Due to the complex nature of this posterior, we resort to approximation to compute the posterior.
- We choose a factorized approximation, $q(\boldsymbol{x})$, to the posterior and we tune its parameters to minimize:

$$\mathrm{KL}(q || p^*) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p^*(\boldsymbol{x})}.$$

Implications for Autonomous Agents

- Explainability
  - Our hybrid theory of memory permits agents to generate self-explanations and explanations of other agents' behavior.
  - Event memory stores agent's goals, beliefs, and intentions.
  - Self-explanations can be generated by retrieving from leaf-level episodes.
  - Explain another agent's behavior by retrieving schematic representation.
  - Schemas can also be used to plausibly explain how events might have occurred.

Implications for Autonomous Agents

- Temporal persistence
  - Event memory-enabled agents can remember previous interactions with users.
  - Temporal persistence can allow an agent to answer follow-up questions about their behavior.
  - Temporal persistence enables systems to describe events at different levels of detail.
  - Event memory-enabled agents can form a theory of mind of their users.

Future Work

- Finish integrating the work into a cognitive architecture (ICARUS).
- Implement virtual assistant with event memory capabilities.
- Test event memory-enabled agents in Minecraft and Starcraft games.

Conclusions

- Poor temporal persistence and opacity in artificial systems limit their ability to collaborate with humans.
- A memory for events can help remedy these issues by storing an agent's temporal contexts and interactions with its environment.
- We proposed a hybrid theory that combines aspects of the causal and simulationist views.
- Our hybrid theory is a promising avenue of research for creating interactive and explainable agents.