Agents, Cognitive Design Patterns and Agentic LLMs

45th Soar Workshop

5 May 2025

Center for Integrated Cognition

Robert Wray

Contributors: James Kirk, John Laird

2025: “The year of the agent”

Nvidia CEO Says 2025 Is the Year of AI Agents

By Tae Kim Follow

Jan 07, 2025, 5:40 pm EST

Share Resize Reprints

Jensen Huang, chief executive officer of Nvidia, speaks during the CES trade show in Las Vegas on Monday. (BRIDGET BENNETT/BLOOMBERG)

(image: barrons.com)

Nvidia CEO Jensen Huang is optimistic that AI agents will become the next big thing for artificial intelligence.

“AI agents are going to get deployed,” he said on Tuesday at a question-and-answer session with financial analysts at the CES tech trade show in Las Vegas. “I think this year we’re going to see it take off.”

ARTIFICIAL INTELLIGENCE

Anthropic’s chief scientist on 4 ways agents will be even better in 2025

The hottest topic in AI is only just getting started.

By Melissa Heikkilä & Will Douglas Heaven January 11, 2025

(image: MIT Tech Review)

Jan 27, 2025 - Axios Events

Agentic AI is a "big next step" in AI's evolution, SAP CEO says

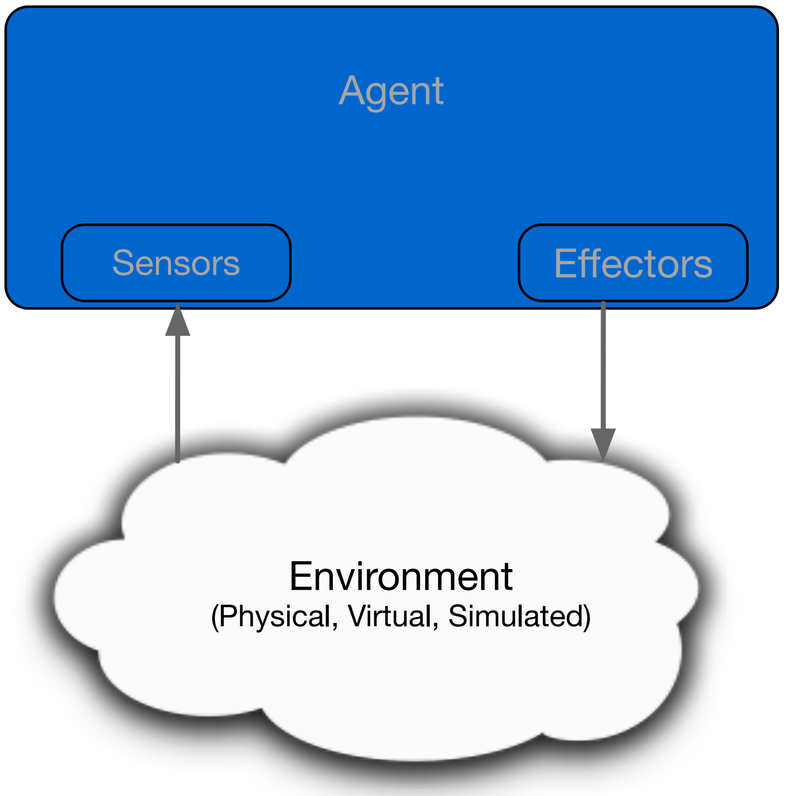
Tyne Phillips Mocek

(image: Axios.com)

What is an agent?

- An agent acts on the environment it inhabits
- A rational agent chooses actions in its (best) interest, given its understanding of the environment (“beliefs”)
  - “Agent” → Intelligent / rational agent
- Properties of rational agents
  - Autonomy
  - Reactivity
  - Proactiveness
  - Social ability

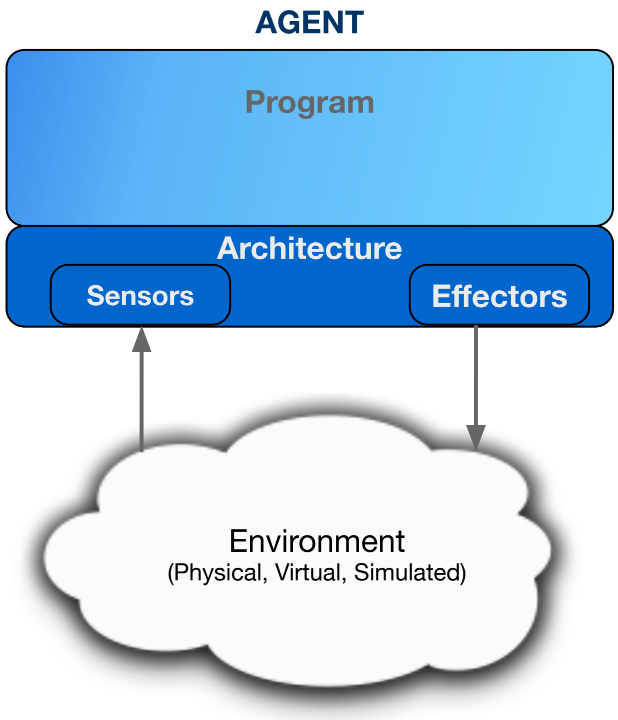
(Wooldridge & Jennings, 1995; Wooldridge, 2000)



What is an Architecture?

- Agent = Architecture + Program (Russell & Norvig, 1995)

- Architectures define/impose fixed constraints on what programs can do (and thus what agents can do)
  - Sensors & Effectors
  - Available memory
  - Processing speed
    - Microcontroller vs. multicore CPU
    - CPU vs. GPU
  - Network connectivity

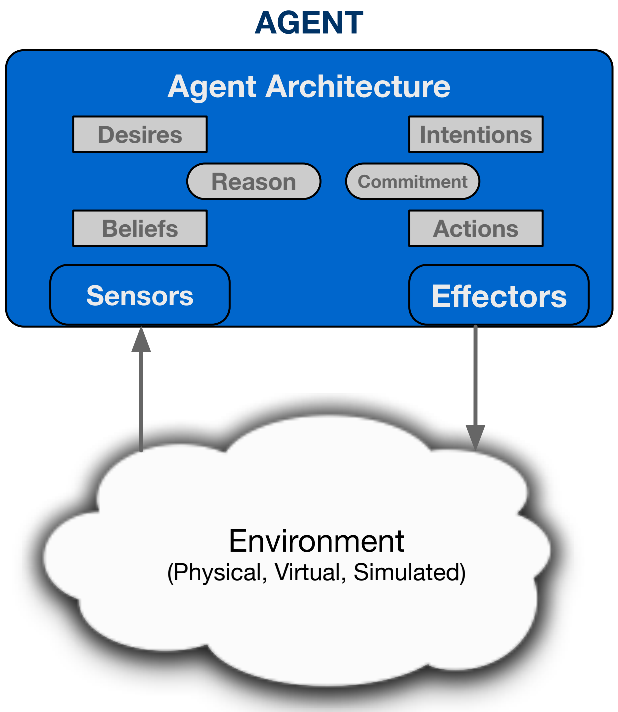


Agent Architectures

- Many “functions” within agent programs recur across tasks
  - Representations & memories, decision processes, learning, …

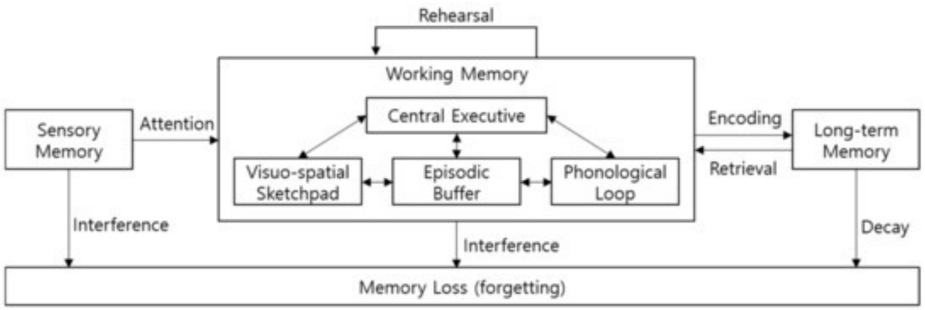
→ Agent architecture: a virtual machine that supports recurring *agent* functions (Jones & Wray, 2004)
  - Task independent implementation

- Example: Belief-Desire-Intention (BDI) Architectures
  - Directly support representation of beliefs, desires, and intentions
  - Consistent processing of BDI primitives
  - dMARS / PRS / UM-PRS / OPRS
  - JAM / JACK
  - Many others … (Silva et. al. 2020)

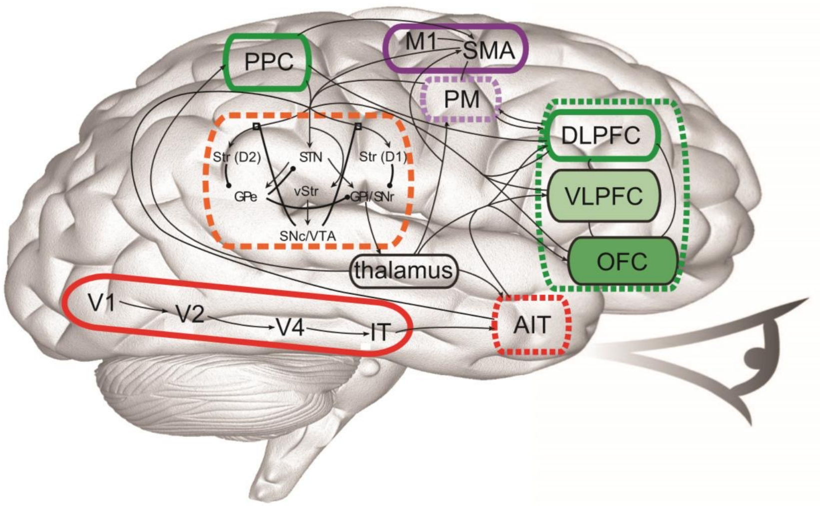


Cognitive Architectures

- Cognitive architecture: Computational theory of the structures and processes of intelligence
- Typically, strongly influenced or inspired by human cognition / mind / brain



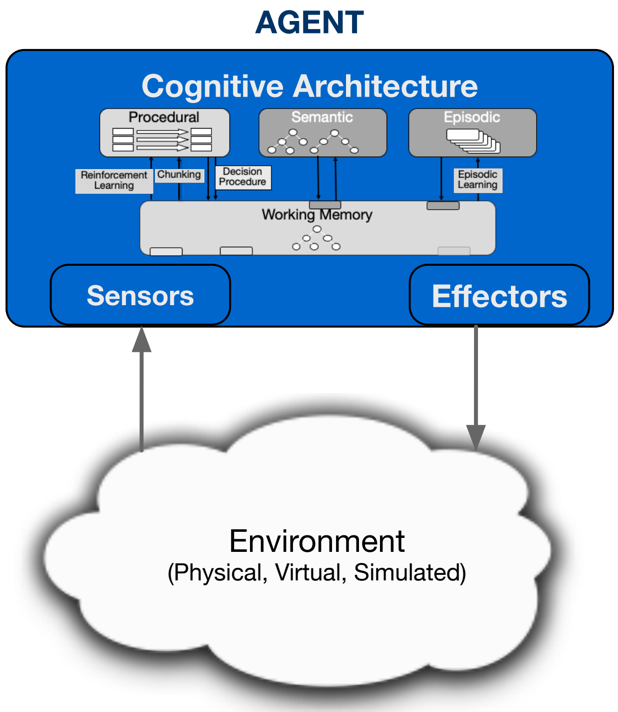
(image: Kang & Bae, 2021)



(image: Eliasmith et al, 2012)

Cognitive Architectures

- Cognitive architecture: Computational theory of the structures and processes of intelligence
- Typically, strongly influenced or inspired by human cognition / mind
- Cognitive architecture: Agent architecture inspired by human cognition (Jones & Wray, 2004) (Wray & Jones, 2004)



Architectures for General Intelligence

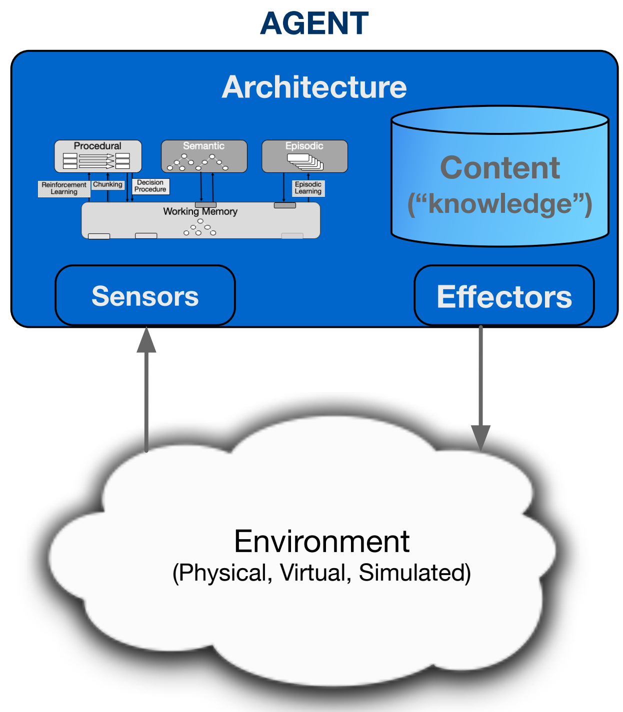
In the ideal case, an agent architecture would be sufficient for any agent task

Reformulation for general intelligence:

Agent = Architecture + Content (aka "knowledge")

- Architecture is generally fixed
- Only content/knowledge differs from task to task

- Recurring functional components (Laird et al, 2017)
- Recurring approaches to primitives and processing (Jones and Wray, 2006; Jones et al, 2009, Silva et al, 2020)

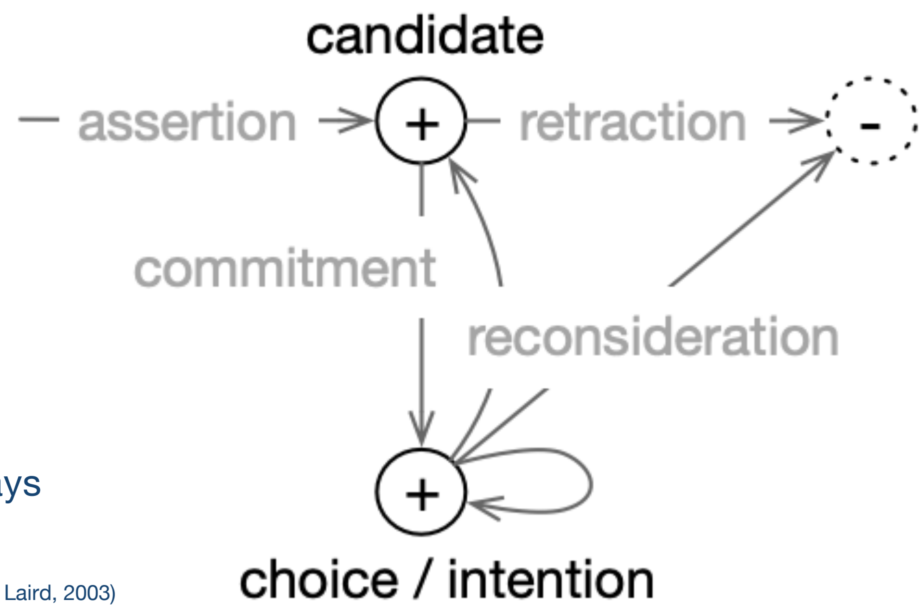


Cognitive Design Patterns

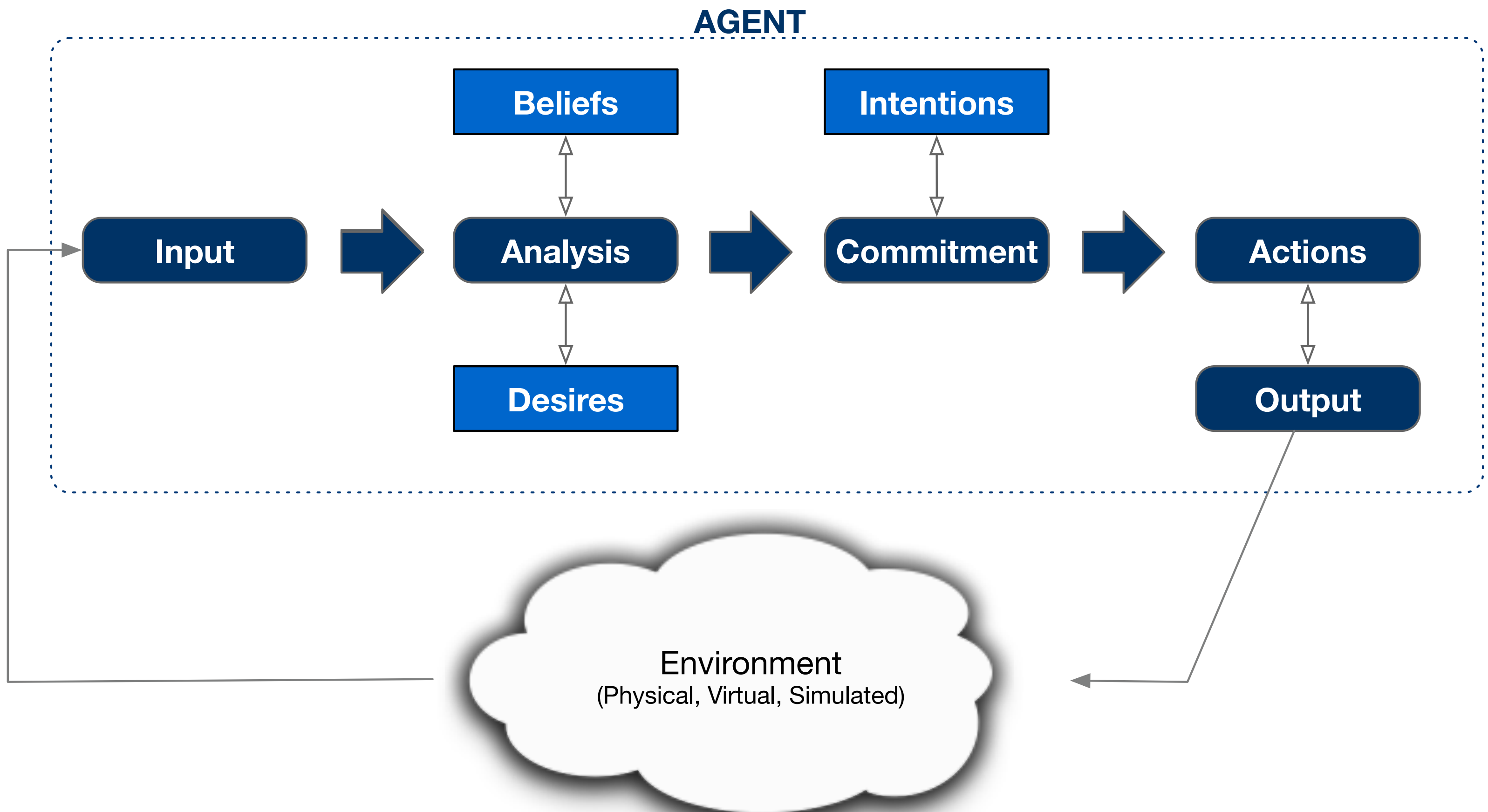
- Software design pattern: Template/description of common solution
- Cognitive design pattern: Abstract description/ specification of computational functionality recurring in cognitive systems
  - Abstract: Not an implementation or (generally) an algorithm
  - Recurring: Occurs in more than an individual architecture or approach

Example:
3-stage Commitment (with Reconsideration)

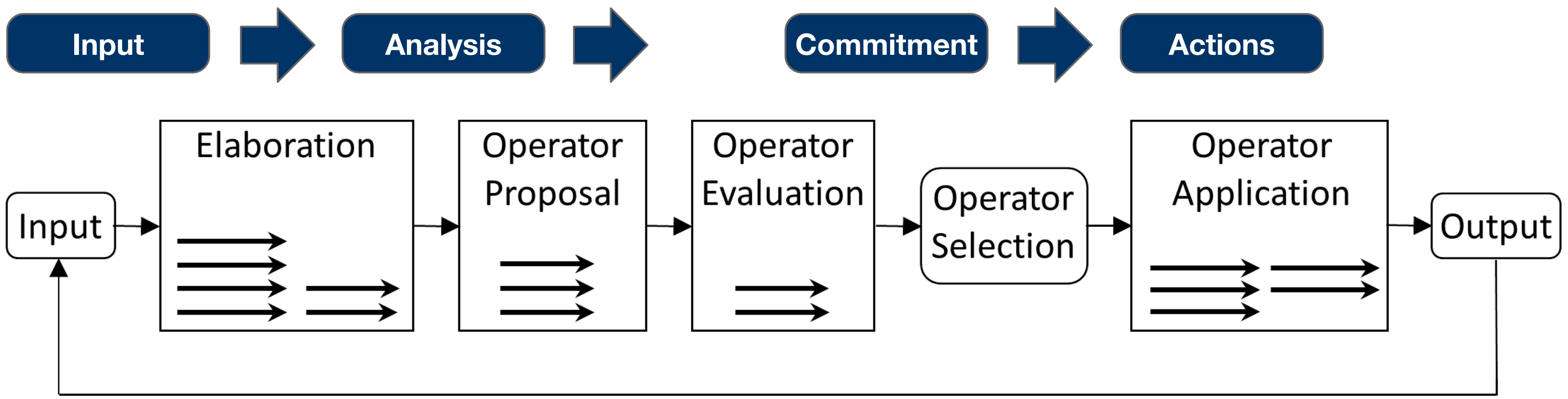
- Pattern: 3 steps for many decisions
  - Generate candidates (assertion)
  - Commit to a candidate (commitment)
  - Determine whether to continue (Reconsideration)
- Recurring:
  - BDI: Desire/Intention
  - Soar: Operator Proposal/Decision
- Abstract:
  - Reconsideration implemented in different ways
  - Decision-theoretic calculus in BDI (Schut et al, 2004)
  - GDS (assumption-based TMS) in Soar (Wray and Laird, 2003)



Example Pattern: BDI Decision Loop



Example Process Pattern: BDI Decision Loop Applied to Soar



Soar Decision Cycle

Cognitive Design Patterns

Cognitive Design Pattern	Examples
Observe-decide-act	BDI: analyze, commit, execute Soar: elaborate/propose, decide, apply (operators)
3-stage memory commitment	BDI: desire, intention, intention reconsideration Soar: operator proposal, selection, retraction Soar: elaboration, instantiation, JTMS reconsideration
Hierarchical decomposition	BDI: hierarchical task networks (HTNs) Soar: operator no-change impasses
Short-term (context) memory	ACT-R: buffers (goal, retrieval, visual, manual, ...) Soar: working memory
Ahistorical KR/memory Retrieval:	ACT-R, Soar: semantic memory ACT-R, Soar: activation-mediated association
Historical KR/memory	Soar: episodic memory
Procedural KR/memory Retrieval: Learning:	ACT-R, Soar: productions, BDI: plans ACT-R, Soar: associative production-condition match ACT-R, Soar: knowledge compilation/chunking

Cognitive Design Patterns & Agentic LLMS

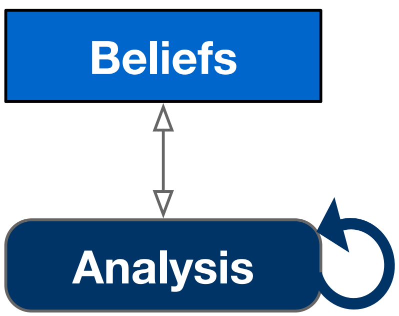
How can cognitive design patterns help us understand, evaluate and predict the future of Agentic LLM research?

1. Analyze existing approaches via established cognitive design patterns
2. Predict functional gaps that might be filled by novel instantiations of cognitive design patterns
3. Identify novel patterns that are made more evident/acute by the properties of LLM computation

Processing Patterns with Agentic LLMs: CoT

Chain of Thought ("think step-by-step") (Wei et al, 2022)

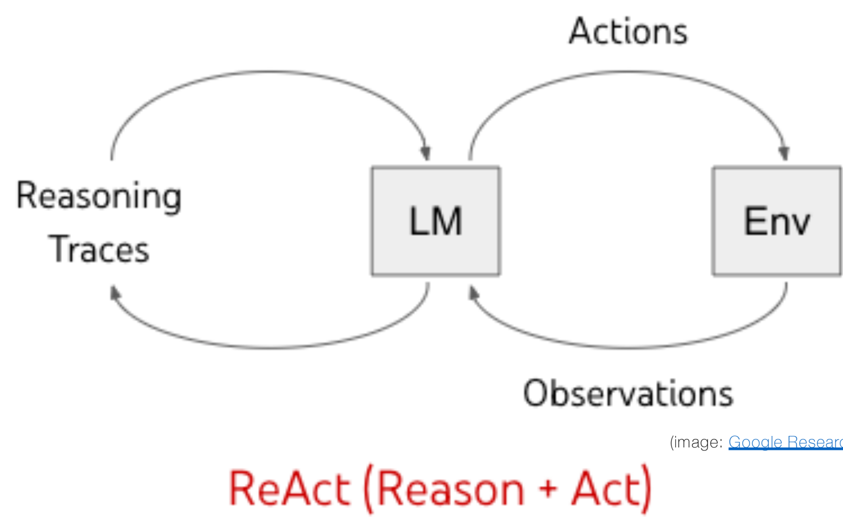
- Makes the implicit explicit via iterative "steps" (functional role comparable to elaboration/belief generation)
- Now incorporated into training process (o1/o3, DeepSeek)



Processing Patterns with Agentic LLMs: ReAct

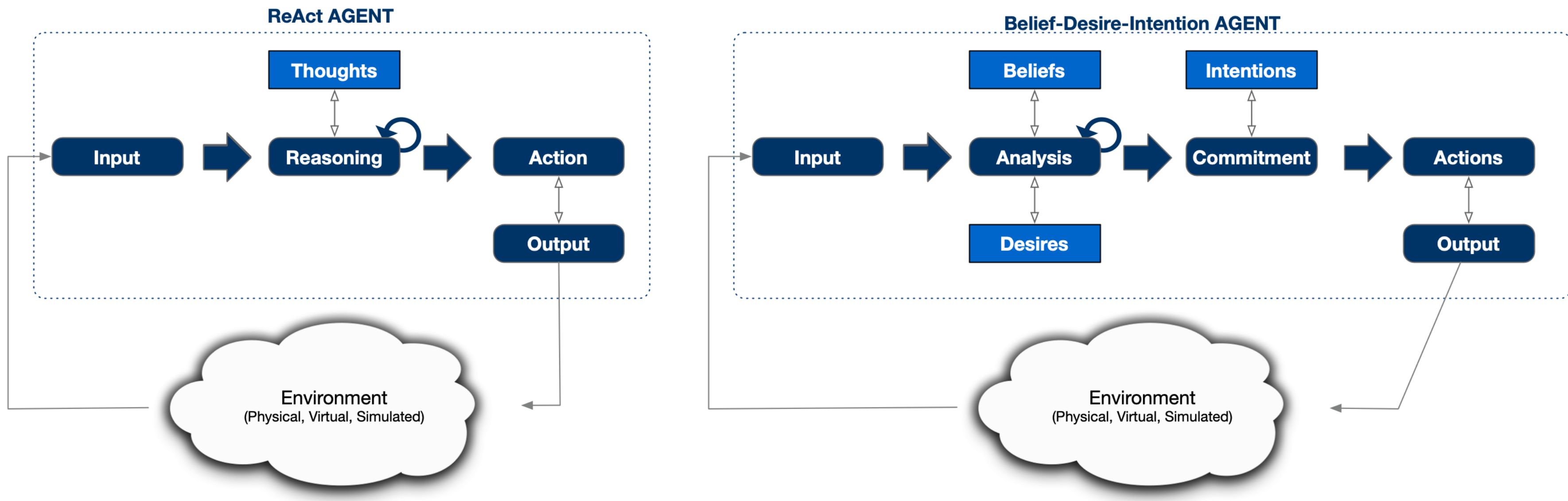
ReAct (Reason and Act) (Yao et al, 2023)

- Splits generation of “thoughts” (internal reasoning) from action
  - Thought ~ Belief/Elaboration
  - Action ~ Intention/Operator Application
- Very influential example of functional decomposition for multi-step reasoning via LLMs



(image: Google Research)

ReAct in Comparison to Observe-Decide-Act



ReAct exhibits 2-stage, not 3-stage commitment

- Does not distinguish intentions from beliefs (all internal assertions are *thoughts*)
- Does not make explicit commitments (and thus no reconsideration)
- *Hypothesis: Explicit commitment and reconsideration would improve ReAct*

Example: Episodic Memory

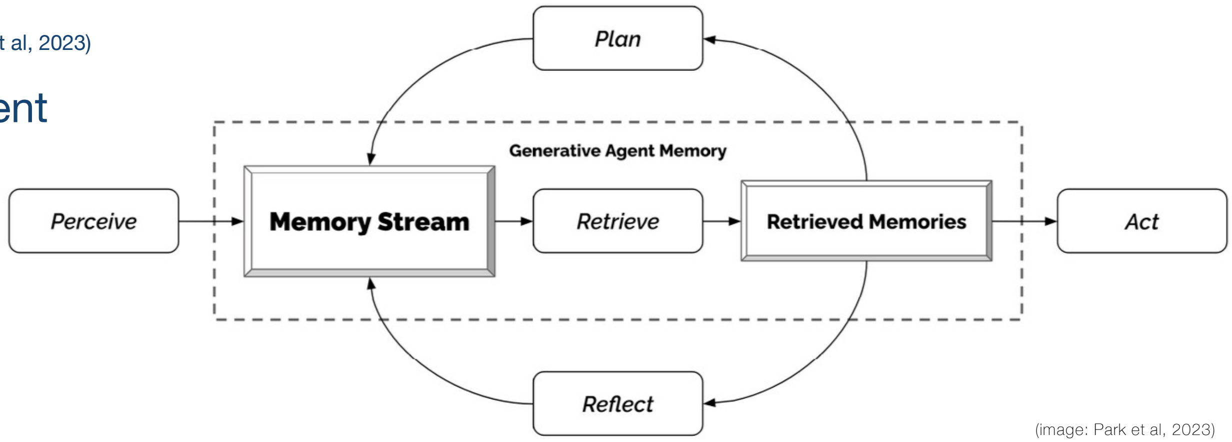
- Episodic memory: Memory of past experiences
  - Long-term (history)
  - Rich in context (distinct from fact memory)
  - Segmentation of events (“episodes”)
  - Automatic, one-shot (no decision to “save” an episode)

- Computational characteristics
  - Efficient storage (compression, forgetting, consolidation)
  - Cue/feature-based, associative retrieval of episodes
    - High correspondence between cues and an episodic aids recall (**Encoding specificity principle**)
    - Retrieval can be deliberate or spontaneous

Episodic Memory in Agentic LLMs

Example: Generative Agents (Park et al, 2023)

- Domain: “Sims”-like environment
- Automatically stores “observations”
- Retrieves memories based on:
  - Recency
  - Importance (agent subjective)
  - Relevance (embedding based)



(image: Park et al, 2023)

- Retrieved memories provide additional context for character planning and action
- Also implements a proto-autobiographical memory (“reflection”) (roughly: “what do I think is important to remember?”)

Comparing Generative Agents & Episodic Mem. Features

Characteristic	In [29]?	Comments
Learning Process		
Automatic	Semi	A periodic process was used to evaluate memories and save a subset.
Autobiographical	Yes	Memories are recorded in the 3rd-person but agents understand memories as being about themselves.
Autonoetic	?	Unclear how agents distinguish current understanding from memory of past experience.
Episodic segmentation	No	Observation span and reflection prompts are based on predefined, static periods (e.g., every 100 observations triggers a reflection).
Variable Length	Yes	Reflections can span variable lengths of agent experience.
Retrieval Process		
Cue-based	Partial	Objects present in agent situation used as retrieval cues
Spontaneous?	Yes	Retrieval is implemented as recurring, automated process
Deliberate?	No	Agents cannot deliberately attempt to construct cues or retrieve memories
Encoding specificity	Partial	Relevance (one of three retrieval criteria) uses semantic similarity, not encoding specificity.

Comparing Generative Agents & Episodic Mem. Features

Characteristic	In [29]?	Comments
Learning Process		
Automatic	**Semi**	**A periodic process was used to evaluate memories and save a subset.**
Autobiographical	Yes	Memories are … stand memori…
Autonoetic	?	Unclear how a… memory of pa…
Episodic segmentation	**No**	**Observation sp… defined, static periods (e.g., every 100 observations triggers a reflection).**
Variable Length	Yes	Reflections can span variable lengths of agent experience.
Retrieval Process		
Cue-based	Partial	Objects present in agent situation used as retrieval cues
Spontaneous?	Yes	Retrieval is implemented as recurring, automated process
Deliberate?	No	Agents cannot deliberately attempt to construct cues or retrieve memories
Encoding specificity	Partial	Relevance (one of three retrieval criteria) uses semantic similarity, not encoding specificity.

Hypothesis: Triggering reflections based on segmented episodes would result in more coherent and readily retrievable summaries of past experience.

Extensive Explorations of Memories for LLMs

Specialized representations

- Annotated, context-extended encoding (Xu et at, 2025)
- Summarized episodes (Park et al, 2023; Zhong et al, 2024)
  - GenAI/LLM approach to episode compression?
- Structured representations (Modarressi et al, 2025)
- Distributed, hierarchical representations (Das et al, 2024)

Encoding/retrieval strategies

- Various properties of transformer architecture / parameters
- Dynamic, hierarchical memories (Hu et al, 2024)
- Task-independent event segmentation (e.g., surprise in Fountas et al, 2024)
- Analogical continuity (Fountas et al, 2024)
- Forgetting (Zhong et al 2024)
- Agent-mediated encode/decode (Xu et al, 2025)
- Constructed memories (Liu, Yang et al, 2023)

Cognitive Design Patterns & Agentic LLMS

How can cognitive design patterns help us understand, evaluate and predict the future of Agentic LLM research?

1. Analyze existing approaches via established cognitive design patterns

   Observe-Decide-Act
   3-Stage Commitment
   Episodic Memory

2. Predict functional gaps that might be filled by novel instantiations of cognitive design patterns

   Reconsideration
   Episodic segmentation
   …

3. Identify novel patterns that are made more evident/acute by the properties of LLM computation

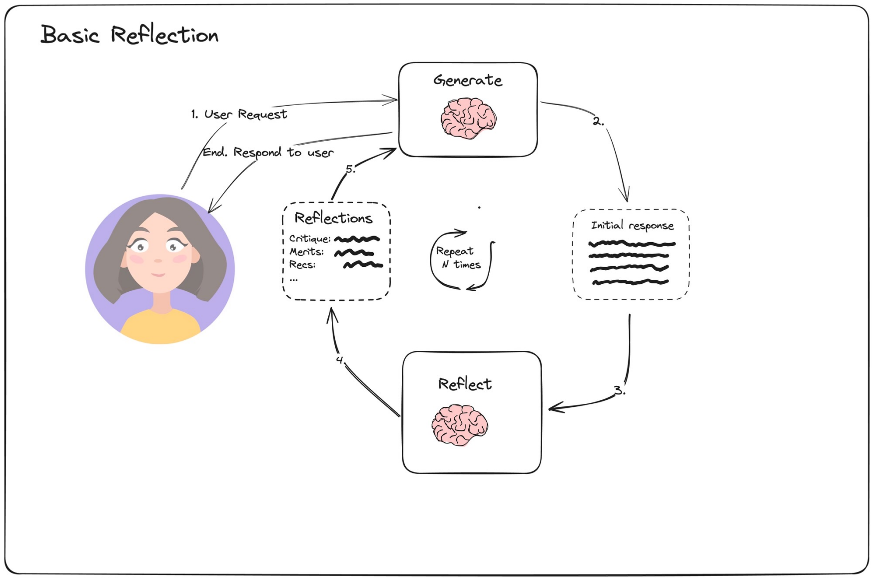
   (Continual) Reflection

New Requirements for Agent LLM Architectures?

- Open Question: Do the computational characteristics of LLMs introduce/require different/novel functional patterns for cognitive behavior?

- Obvious issues: Hallucination, reasoning missteps, ignoring rules, …
  - BDI and cognitive arch. agents generally do not exhibit any of these issues

- Self-correction is one approach to mitigating these issues in LLMs
  - Use an LLM to critique its (or another LLM's) previous outputs (Shinn et al, 2023; Pan et al, 2023)
  - Reflexion-style self-critique incorporated into LangGraph



(image: LangChain)

Conclusions

Nuggets

- Cognitive design patterns are useful as a technique for understanding similarities and differences in agent/cognitive architectures ("Comparative cognitive systems")
- Cognitive design patterns can help us organize and understand explosion of research in Agentic LLMs
- Analysis leads to specific, testable hypotheses informed by established features of patterns

Coal

- "Human-inspired" has little actual meaning in recent work (very rough analogies)
- Lack of visibility of understanding from agent/cognitive architectures to Agentic LLM communities
- Who would support/fund cognitive-architecture-informed Agentic LLM research?

Acknowledgments

CIC Research results included in this presentation were supported in part by the Office of Naval Research, contract N00014-21-1-2369 and DARPA prime contract HQ072716D0006.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.