Lecture 4.

Cognitive Architectures

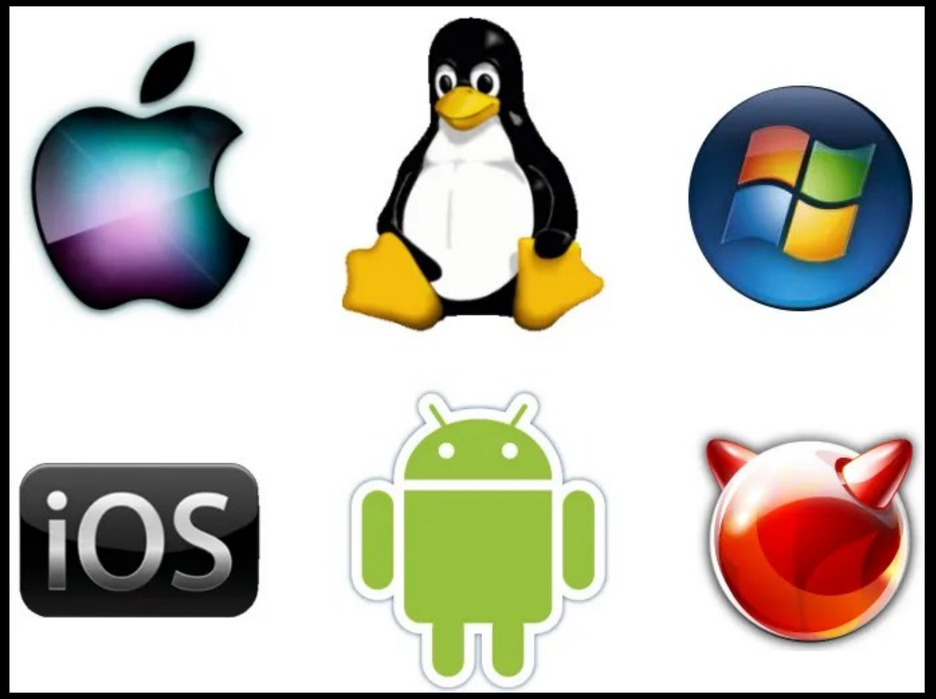
CS 222: AI Agents and Simulations
Stanford University

Joon Sung Park

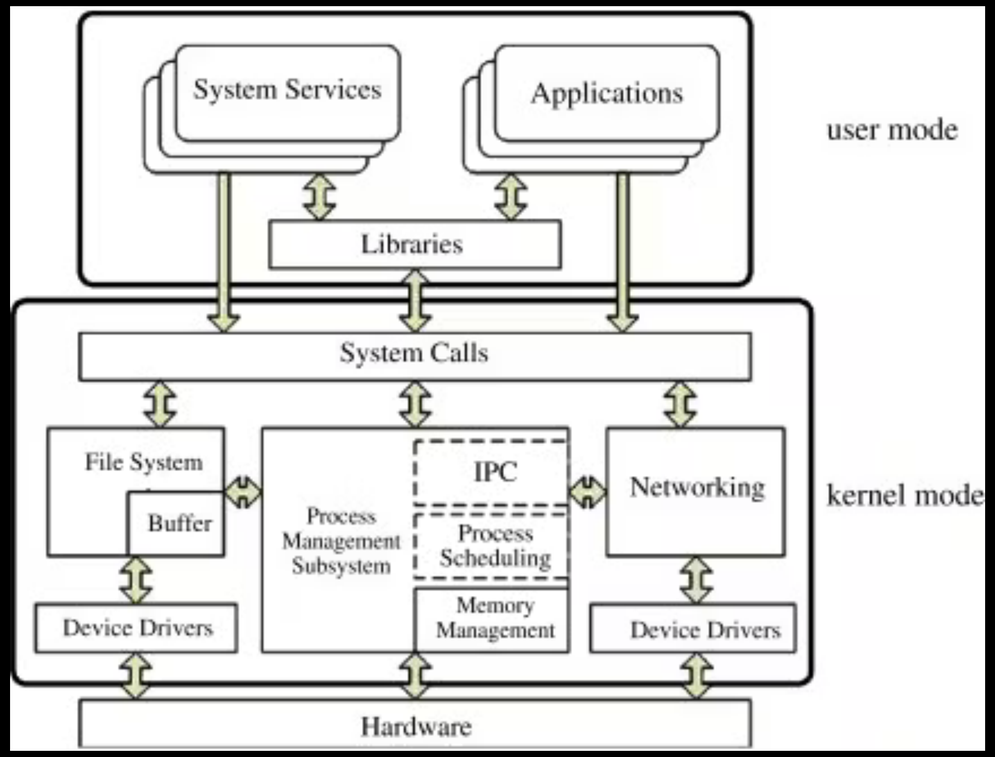
Q. What are “architectures"?

Quick poll:
https://pollev.com/helenav330

Architectures in OS and hardware systems



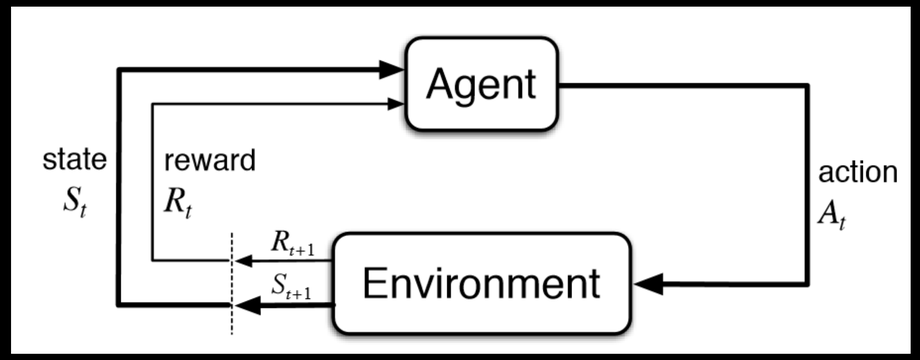
Operating Systems



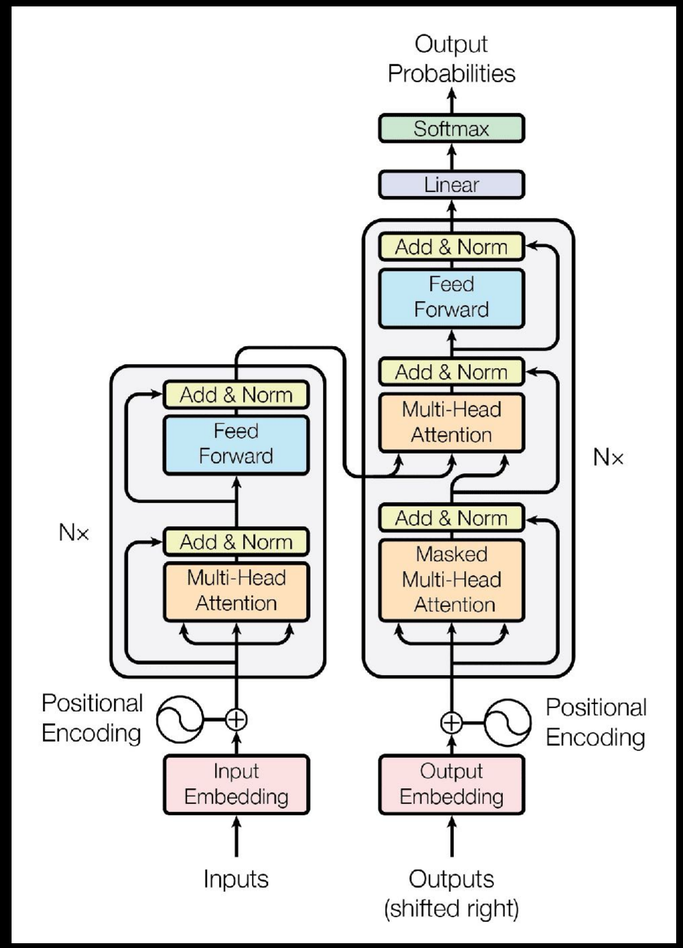
PC Hardware

M. Mathew, Understanding Operating System Architecture: Key Components and Features. Hashnode (2023); https://merwin.hashnode.dev/understanding-operating-system-architecture-key-components-and-features

Architectures in AI



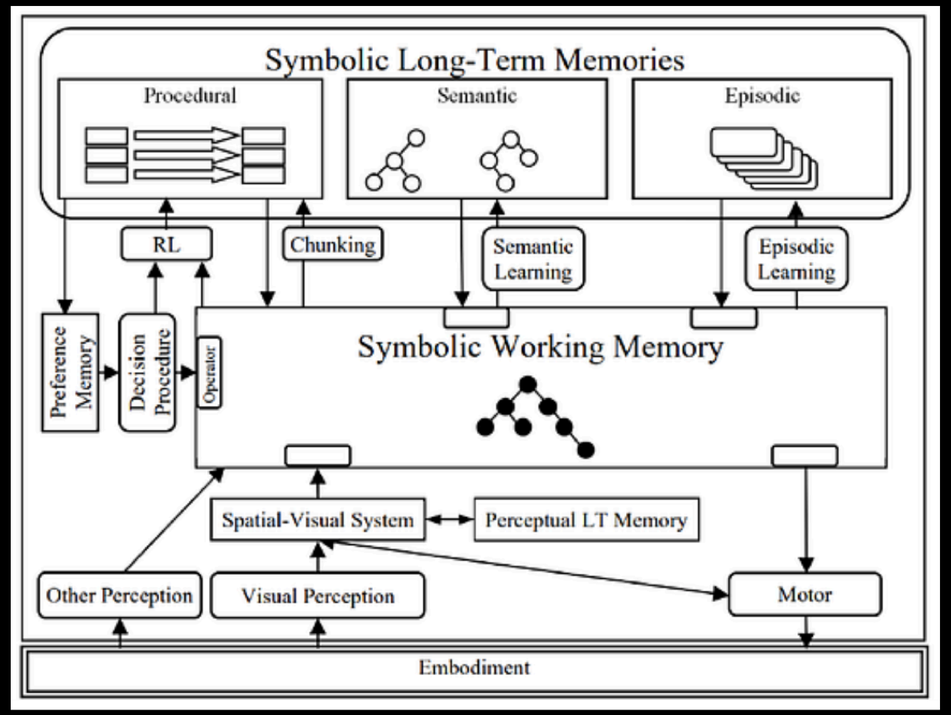
Reinforcement Learning



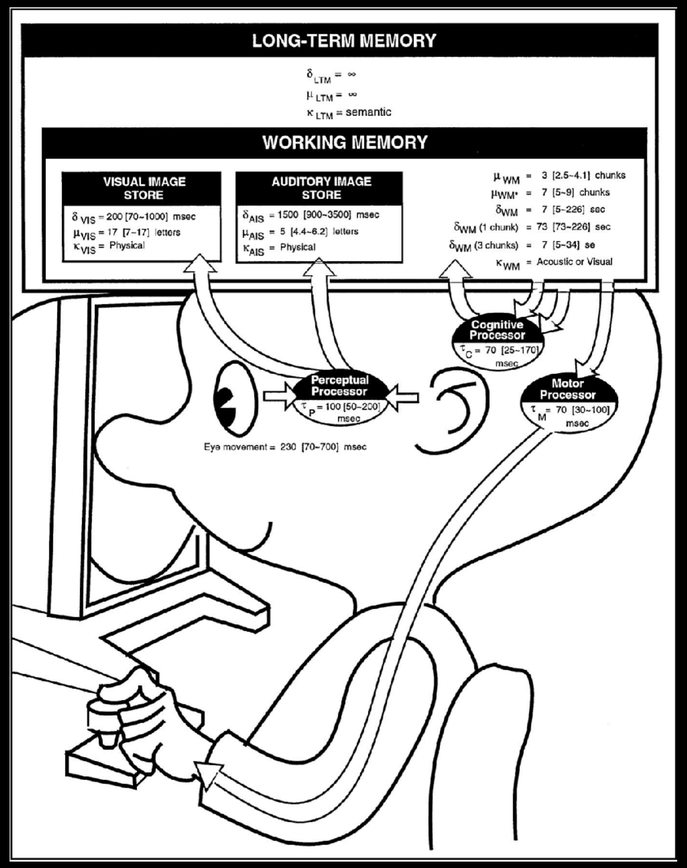
Transformer

R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction (MIT Press, ed. 2, 2018).
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998-6008 (2017).

Cognitive architectures



SOAR



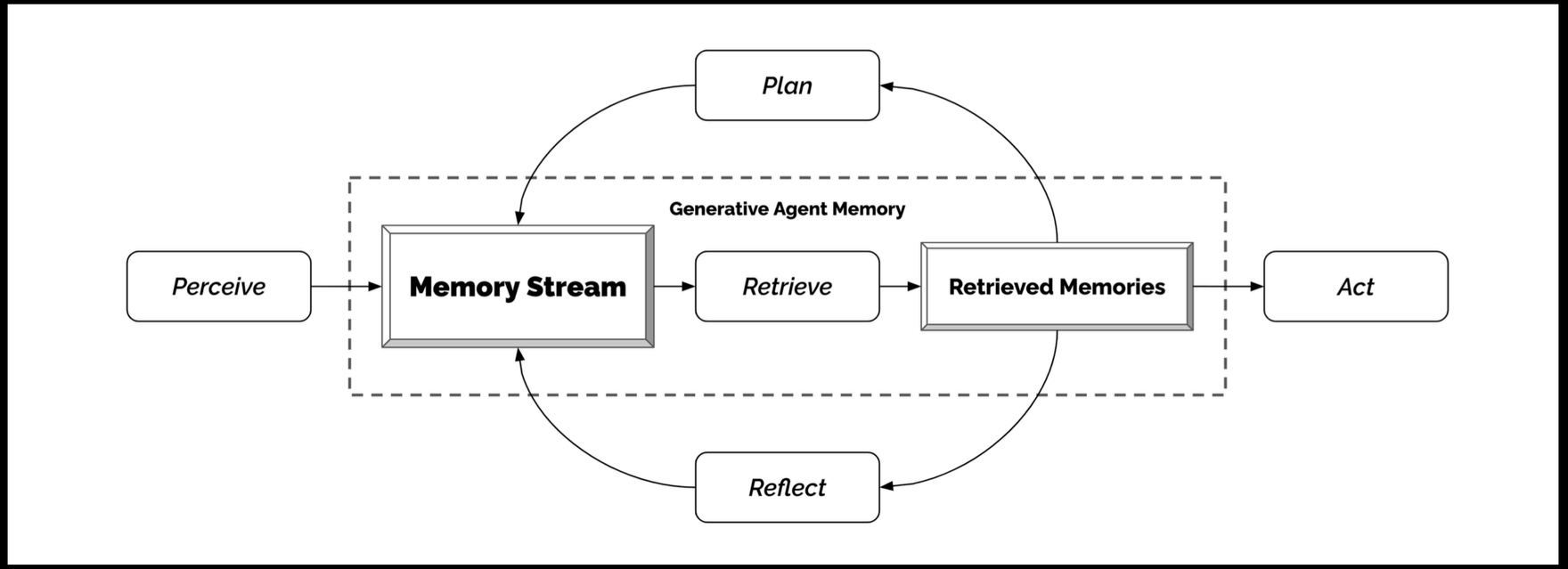
GOMS

J. F. Lehman, et al., A Gentle Introduction to Soar, an Architecture for Human Cognition: 2006 Update.
SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).

Architectures are both a description of a functional system and a theory.

They are not a step-by-step recipe; rather, they offer a perspective on how a system should work.

Today: Cognitive architectures and the architectures of generative agents



Generative Agents

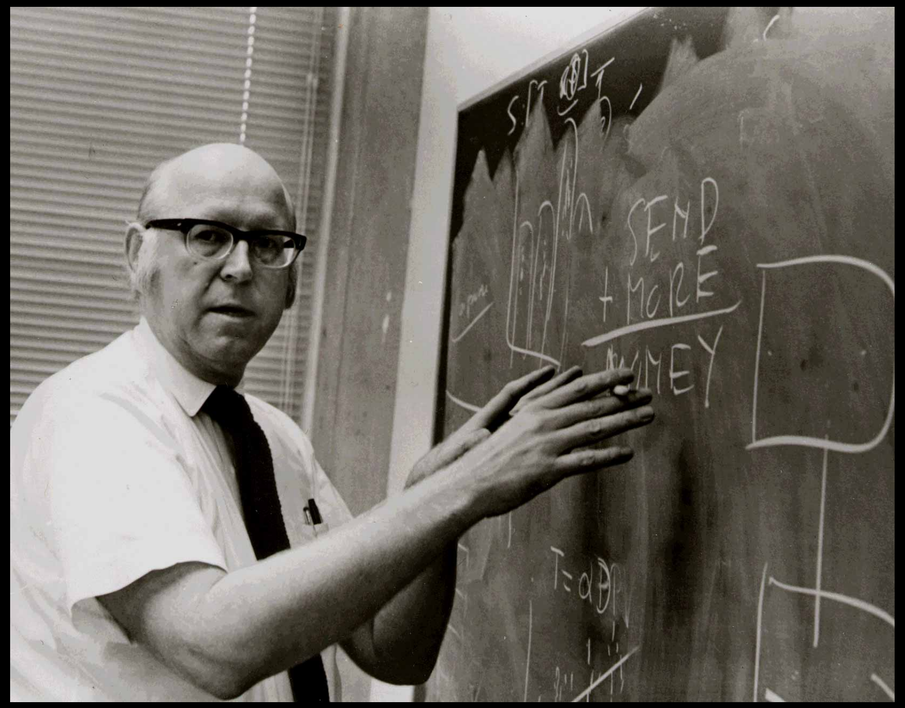
J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023)

A brief history of cognitive science

“Desires and Diversions” lecture

https://www.youtube.com/watch?v=vpfAOBbtGTo

A. Newell, Desires and Diversions (Carnegie Mellon University, Pittsburgh, PA, 1991).



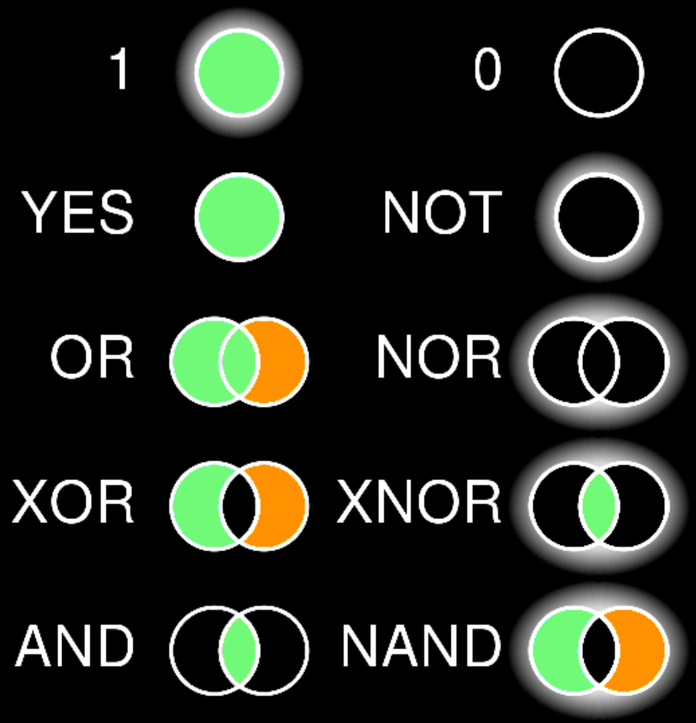
Allen Newell

Bachelor's degree in physics from Stanford in 1949

Graduate degree in mathematics at Princeton in 1949–1950 (exposed to the then "unknown" field of game theory)

Ph.D. from the Tepper School of Business at Carnegie Mellon under Herbert Simon

Logic Theorist (1955)



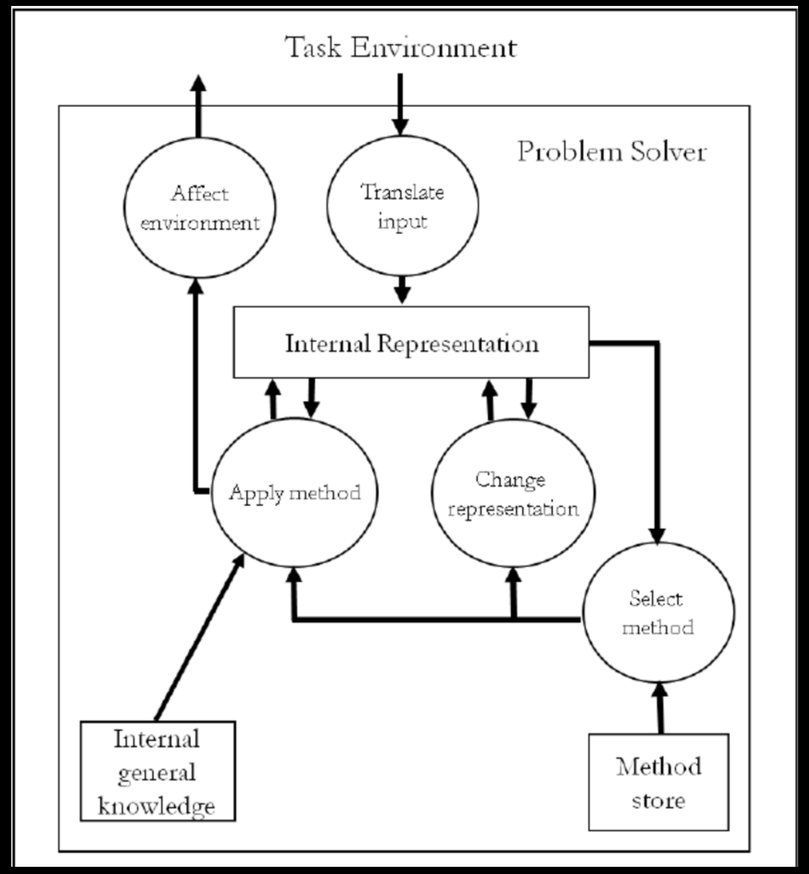
Possibly the first AI program in history.

Written in 1955 by Allen Newell, Herbert A. Simon, and Cliff Shaw, and presented at the Dartmouth workshop.

It proved 38 of the first 52 theorems in chapter two of Whitehead and Bertrand Russell's Principia Mathematica and found new, shorter proofs for some of them.

A. Newell, H. A. Simon, C. Shaw, The Logic Theory Machine. IRE Trans. Inf. Theory 2, 61-79 (1956).

General Problem Solver (1956)

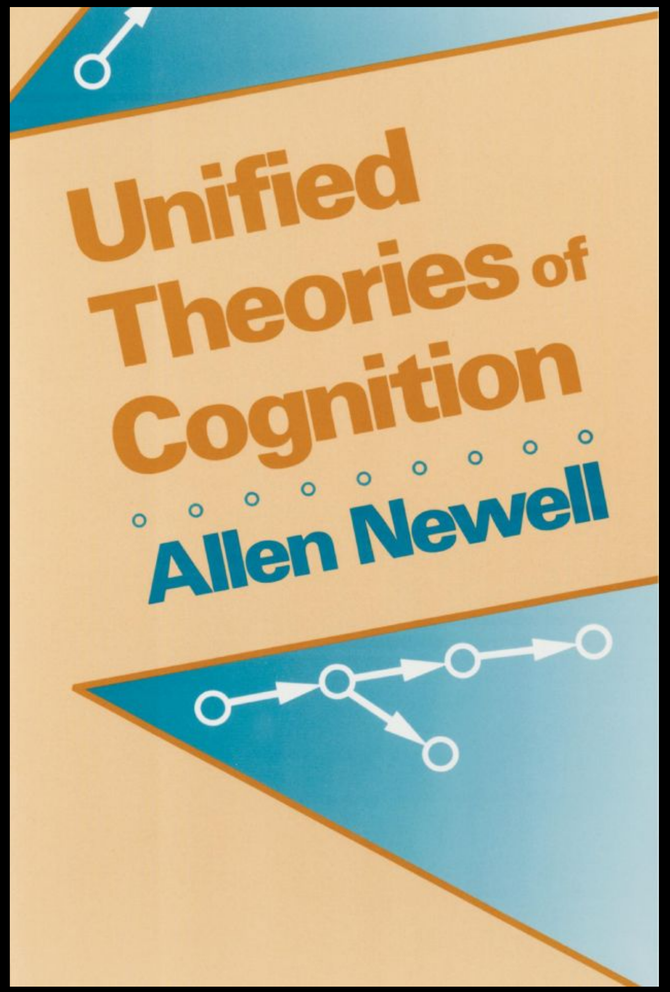


Written in 1956 by Allen Newell, Herbert A. Simon, and Cliff Shaw, and presented at the Dartmouth workshop.

Any problem that can be expressed as a set of well-formed formulas (WFFs) or Horn clauses, and that constitutes a directed graph with one or more sources and sinks (i.e., desired conclusions), can, in principle, be solved by GPS.

A. Newell, J. C. Shaw, H. A. Simon, Report on a general problem-solving program. Proceedings of the International Conference on Information Processing, 256-264 (UNESCO House, Paris, 1959).

Unified Theories of Cognitions



Written in 1990 by Allen Newell.

Newell argues for the need for a set of general assumptions for cognitive models that account for all aspects of cognition: a unified theory of cognition, or cognitive architecture.

A. Newell, Unified Theories of Cognition (Harvard University Press, Cambridge, MA, 1990).

Early observation: Scholars in cognitive psychology began to propose that computers process information similarly to the human mind.

— Can we understand how the human mind works by illustrating it with cognitive architectures?

— Can we create general-purpose computational agents that solve human tasks?

An interesting parallel:

Early observation: Scholars in psychology and AI began to propose that computers process information similarly to the human mind.

Classic cognitive architectures

Early observation: Scholars in HCI and AI began to propose that generative AI encodes and generates human-like behaviors.

Today

Why does history matter to our study?

— The goals set in the early days of a field are often audacious, sometimes premature for the field, but inspiring nonetheless.

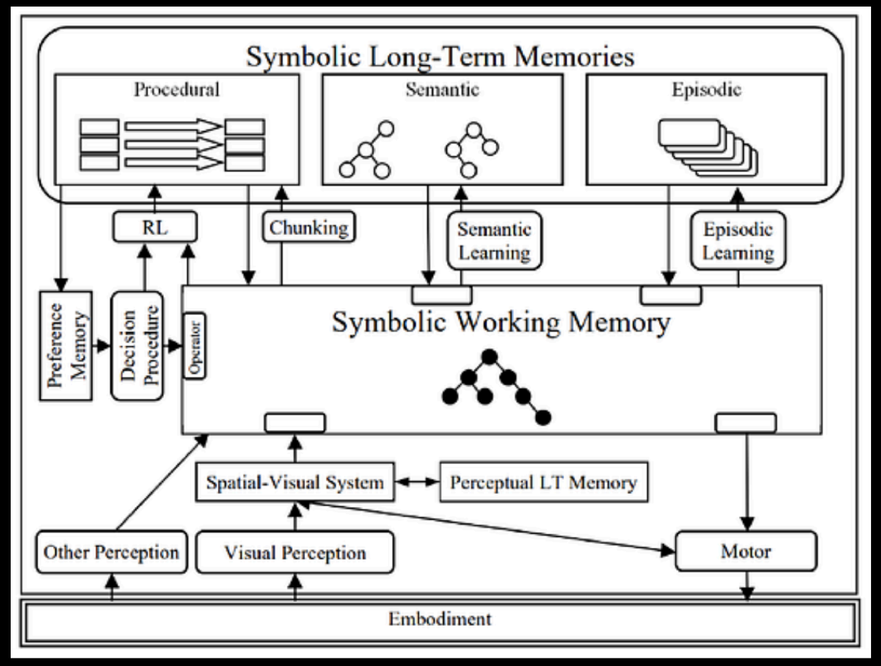
— History often repeats itself and provides us with a useful guide as we navigate and build a new field.

Examples of Cognitive Architectures

Cognitive architecture suggests a theory of how human cognition works.

"... a cognitive architecture as a theory of the fixed mechanisms and structures that underlie human cognition. Factoring out what is common across cognitive behaviors, across the phenomena explained by microtheories, seems to us to be a significant step toward producing a unified theory of cognition..."

SOAR (1983)

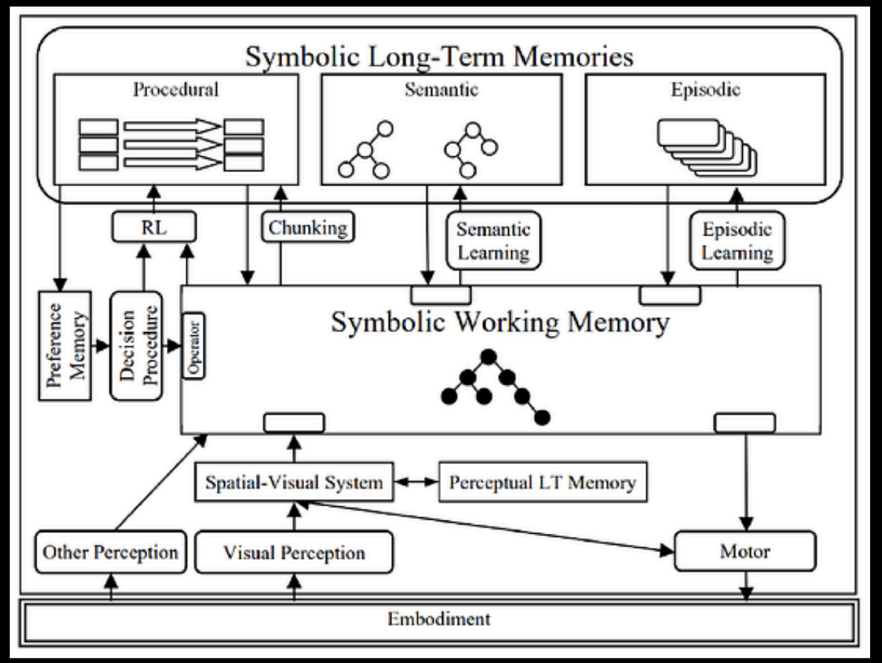


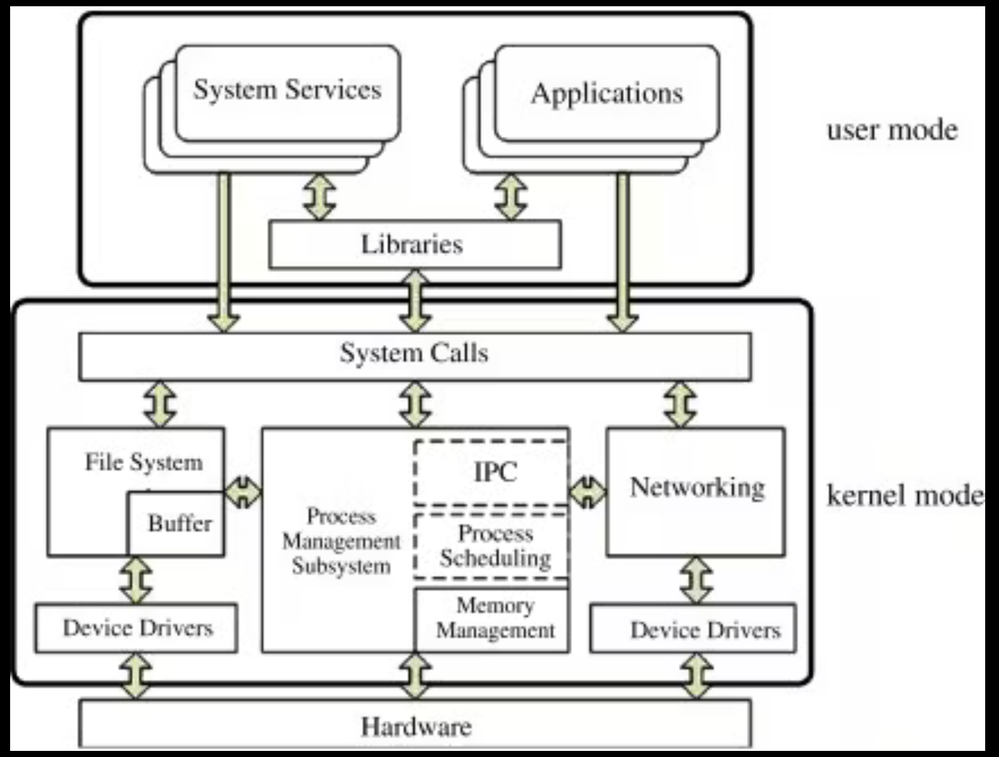
Began as John Laird's PhD thesis (working with Allen Newell, and Paul Rosenbloom)

SOAR posits the Problem Space Hypothesis: All goal-oriented behavior can be cast as search through a space of possible states (a problem space) while attempting to achieve a goal. At each step, a single operator is selected, and then applied to the agent's current state, which can lead to internal changes, such as retrieval of knowledge from long-term memory or modifications or external actions in the world.

J. E. Laird, A. Newell, P. S. Rosenbloom, SOAR: An architecture for general intelligence. Artif. Intell. 33, 1-64 (1987).

The SOAR architecture resembles that of computer architecture.





J. E. Laird, A. Newell, P. S. Rosenbloom, SOAR: An architecture for general intelligence. Artif. Intell. 33, 1-64 (1987).

Example of how the Problem Space Hypothesis manifests

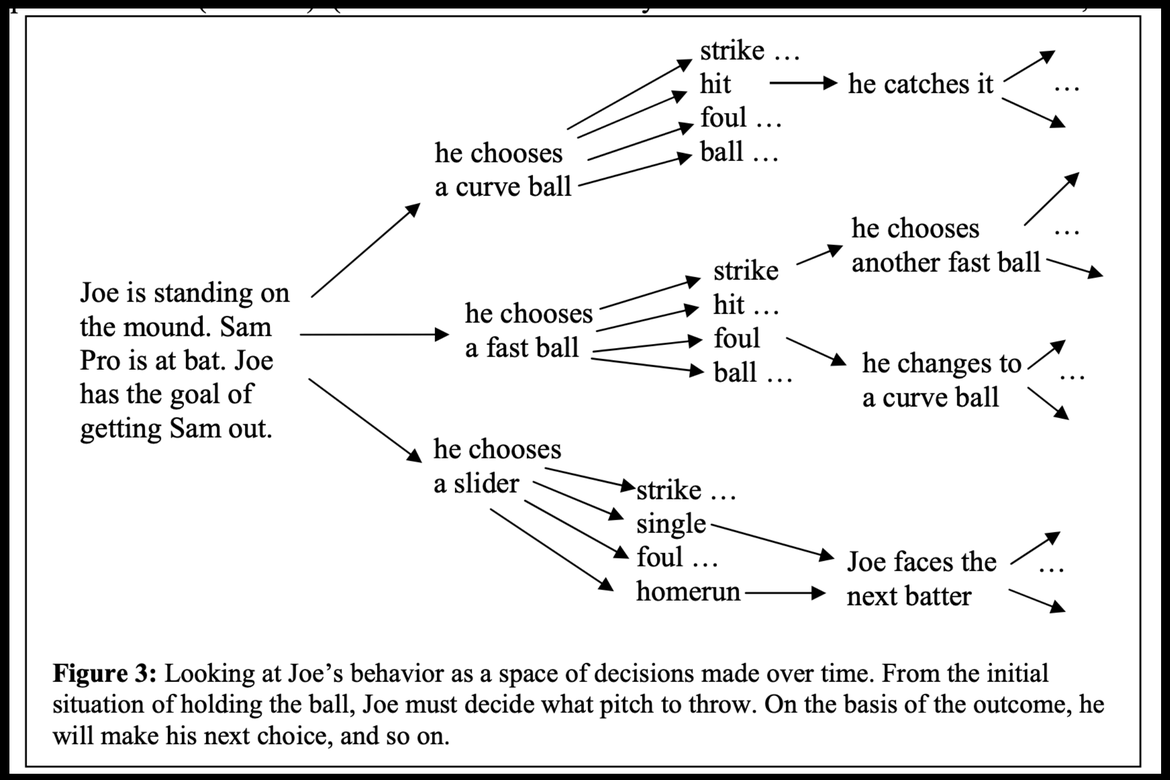


Figure 3: Looking at Joe's behavior as a space of decisions made over time. From the initial situation of holding the ball, Joe must decide what pitch to throw. On the basis of the outcome, he will make his next choice, and so on.

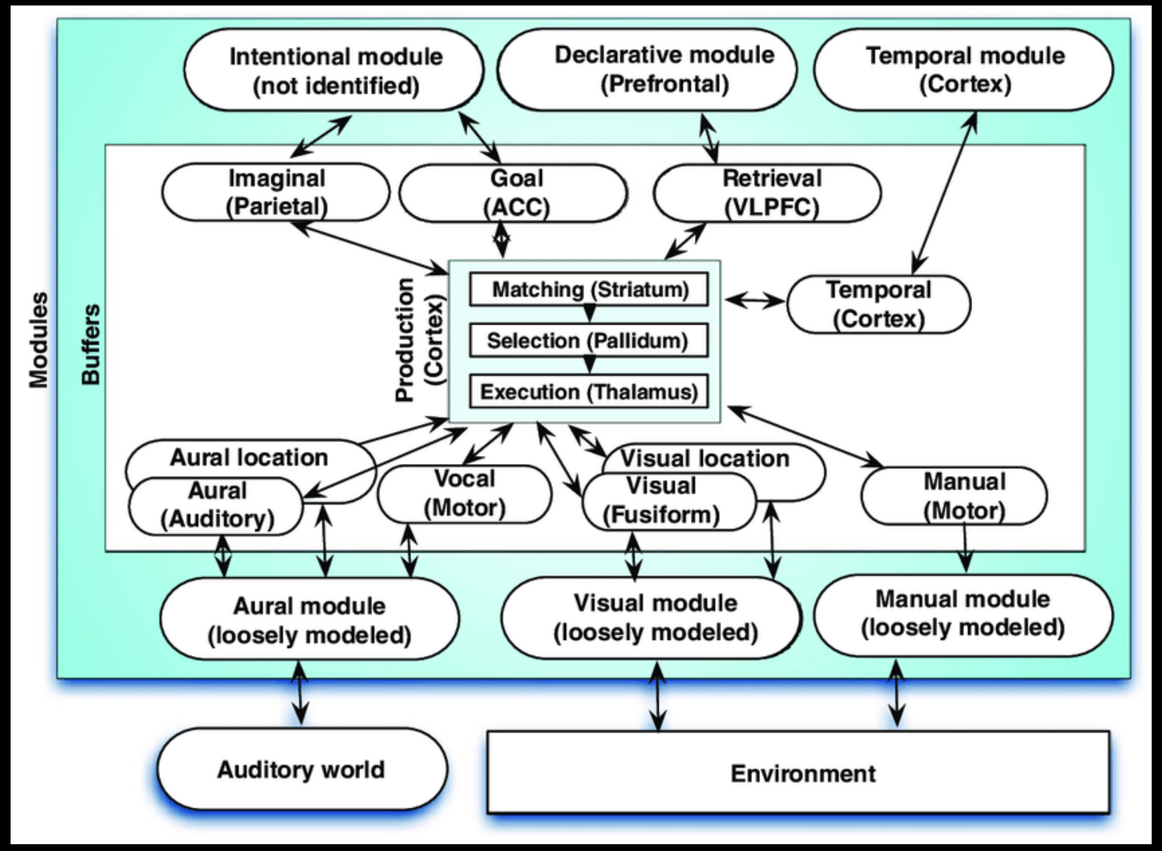
A static view of Joe's life, which we can use to talk about all the possible actions he might take in a particular situation, and a dynamic view of Joe's life, which we can use to talk about the actual path his behavior moves him along.

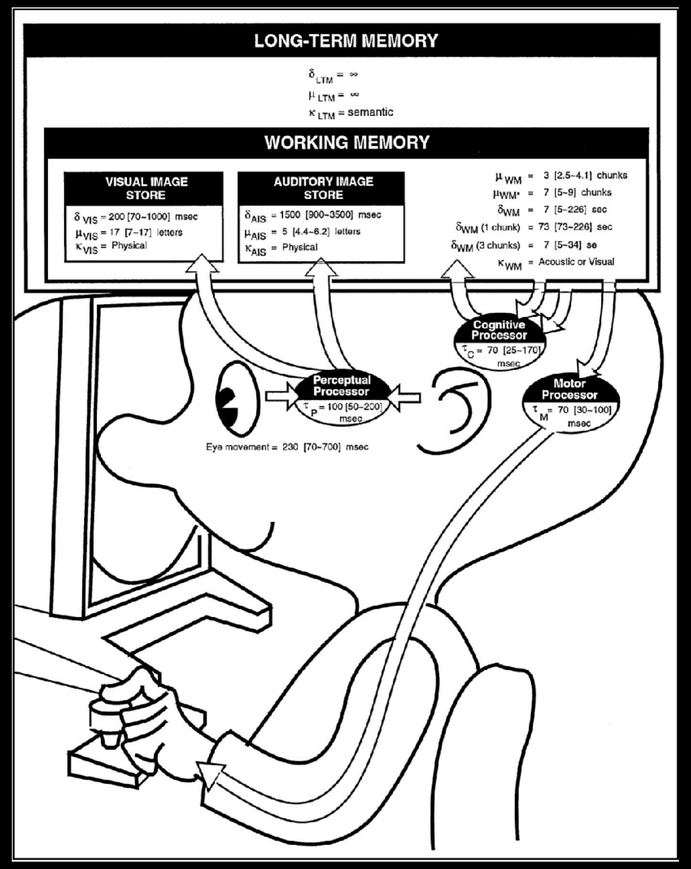
J. E. Laird, A. Newell, P. S. Rosenbloom, SOAR: An architecture for general intelligence. Artif. Intell. 33, 1-64 (1987).

In a way, cognitive architectures are a stylized caricature of human cognition.

“Soar is one theory of what is common to the wide array of behaviors we think of as intelligent. It is not the only such theory (see, e.g., Anderson, 1993; Kieras, Wood, and Meyer 1997, Langley and Laird 2002), but it is the one we will explore in detail.”

Other examples of classic cognitive architectures





ACT-R

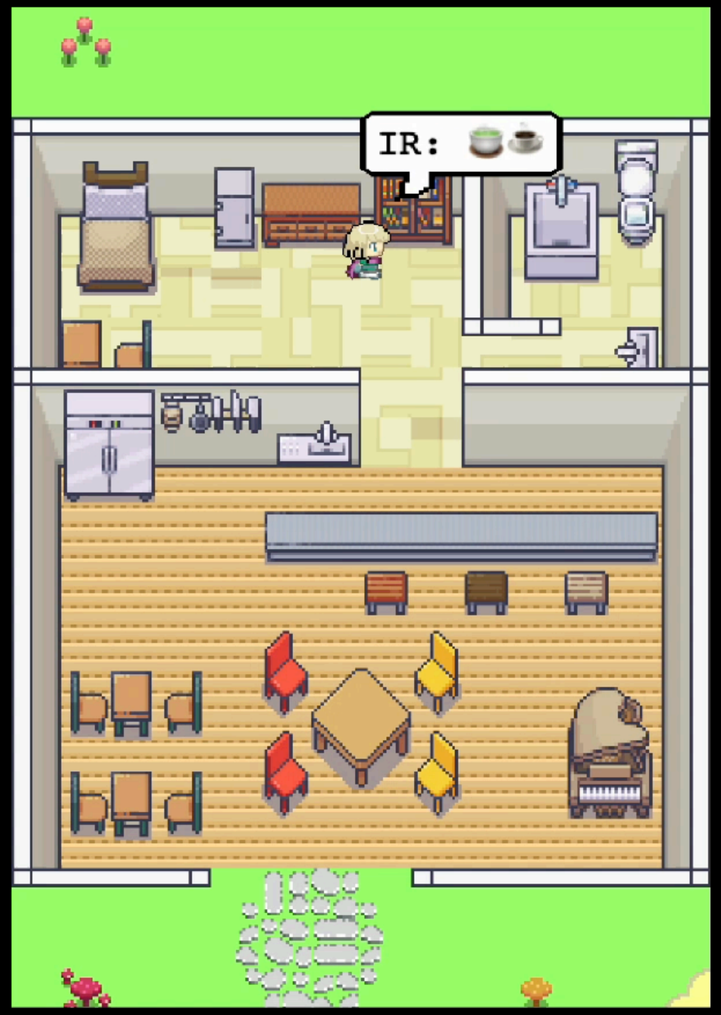
GOMS

J. R. Anderson, C. Lebiere, The Atomic Components of Thought (Lawrence Erlbaum Associates, Mahwah, NJ, 1998).
SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).

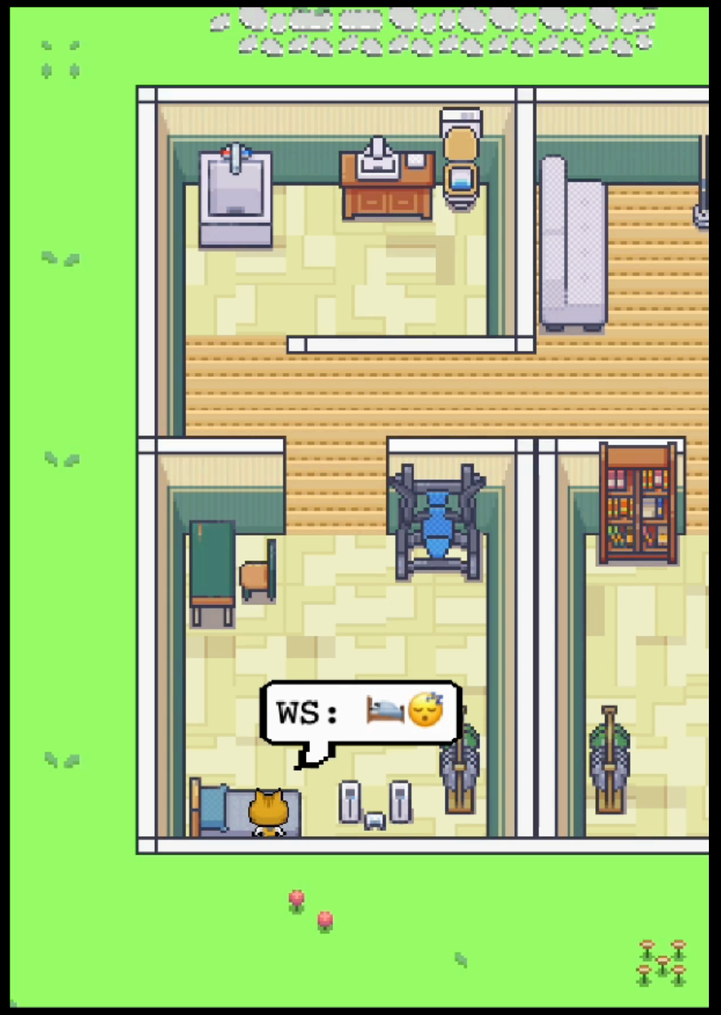
Games and game NPCs have often served as a testbed for cognitive architectures.

Architecture of Generative Agents

Old couple
shares a routine



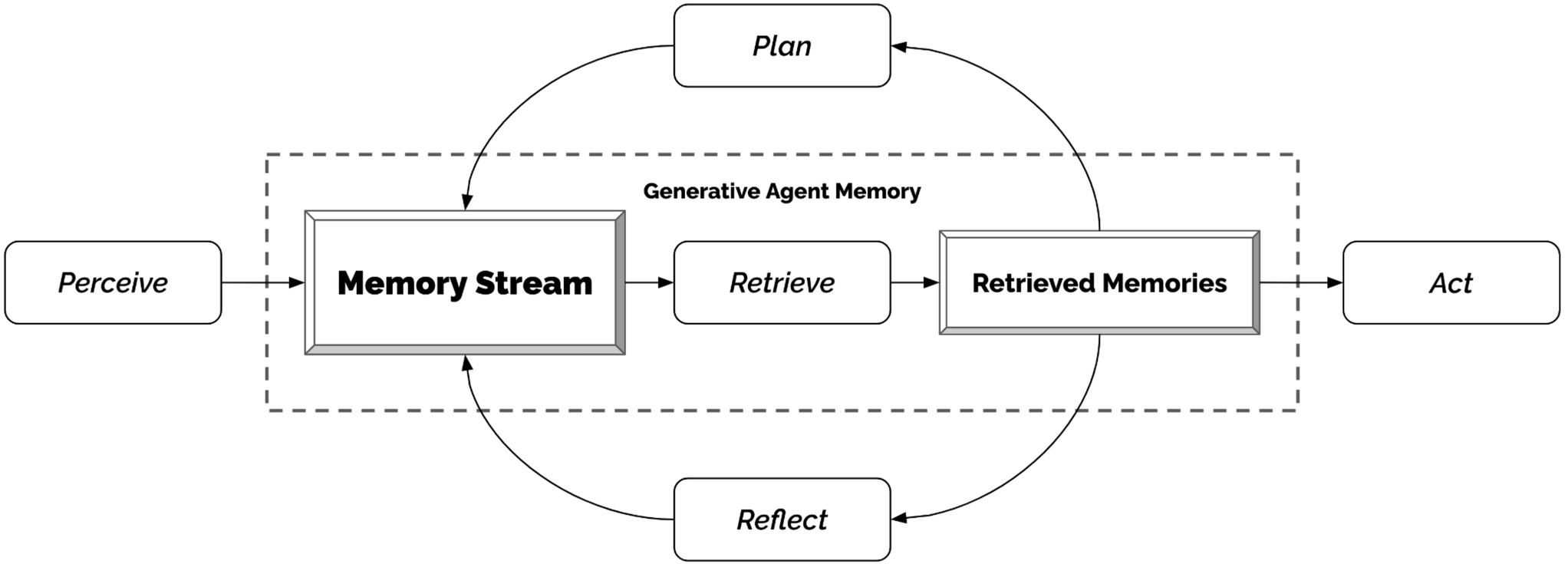
Cafe owner
opens the shop



Student athlete
goes for a run

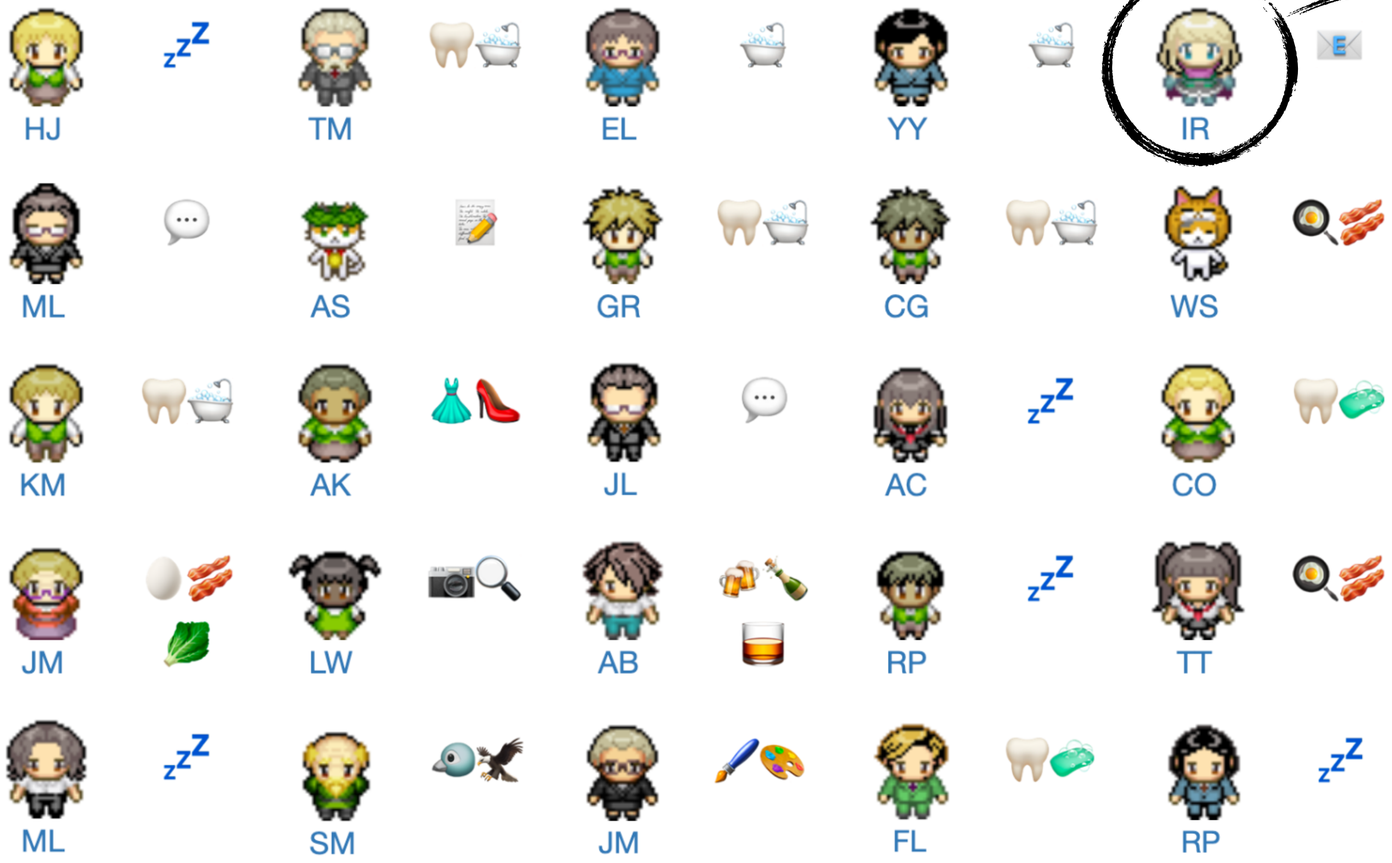

LW:
RP:
AC:
FL:
HJ:
AB:
RP:
IR:
GR:
CG:
WS:
SM:
JL:
KM:
AK:
ML:
AS:
YY:
JM:
TT:
CO:
JM:
TM:
ML:
EL:

Generative Agents



Smallville is a game world inhabited by generative agents

25 agents populate Smallville

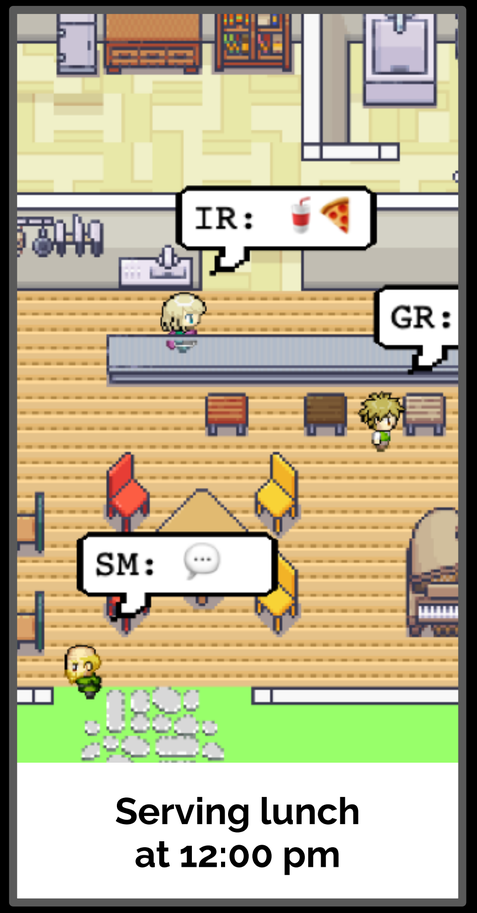


"Isabella Rodriguez is the owner of Hobbs Cafe who loves to make people feel welcome; [...] Isabella Rodriguez is planning on having a Valentine's Day party at Hobbs Cafe at 5pm."

Agents plan and execute their daily behaviors

Washing up
at 6:05 am

Opening cafe
at 8:00 am



Serving lunch
at 12:00 pm

Buying supplies
at 6:25 pm

Agents' actions impact the game environment

"Make and drink coffee"

Cup -> cleaned

Coffee machine
-> turned on

Chair -> occupied

Agents remember their interactions

While taking a walk, Sam meets Latoya, and they introduce themselves:

`[Latoya]: I'm here to take some photos for a project I'm working on.`

Agents remember their interactions

While taking a walk, Sam meets Latoya, and they introduce themselves:

[Latoya]: I'm here to take some photos for a project I'm working on.

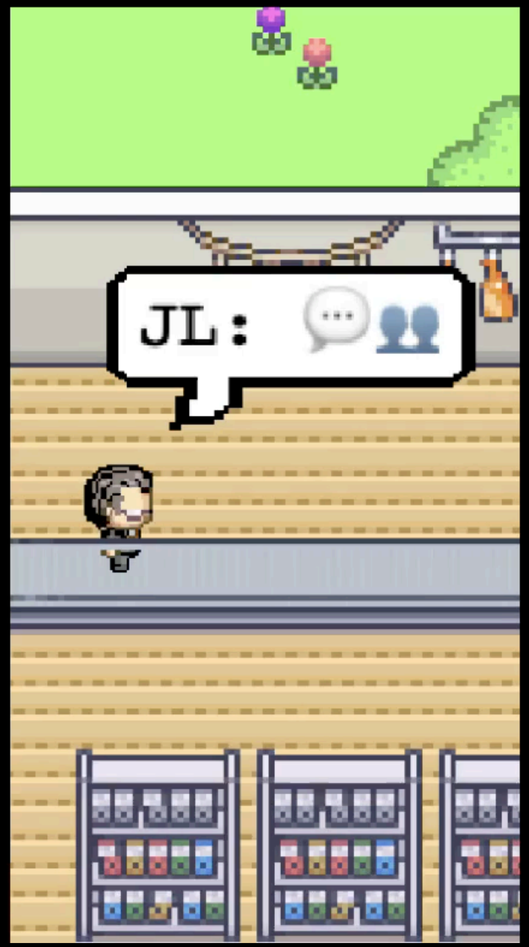
The next day...

[Sam]: Hi, Latoya. How is your project going?

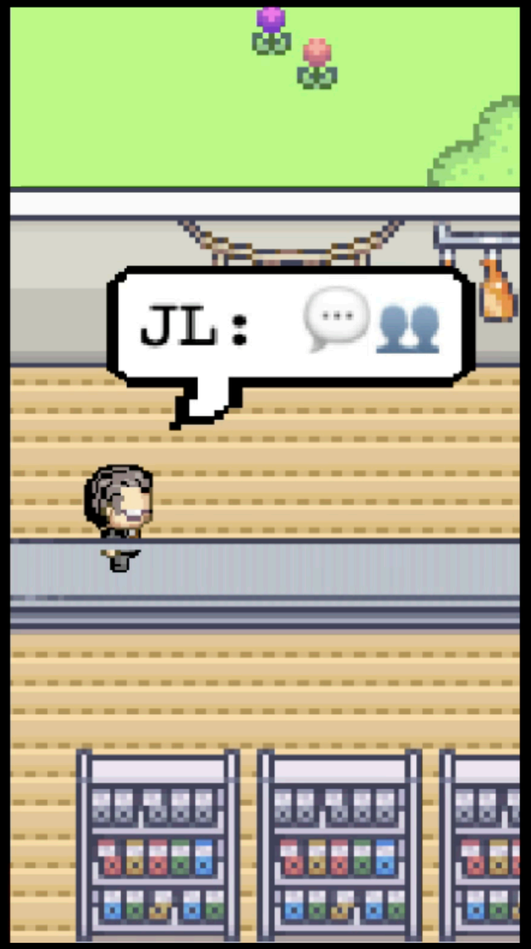
[Latoya]: Hi, Sam. It's going well!

You can interact with Smallville


WS:
ML:

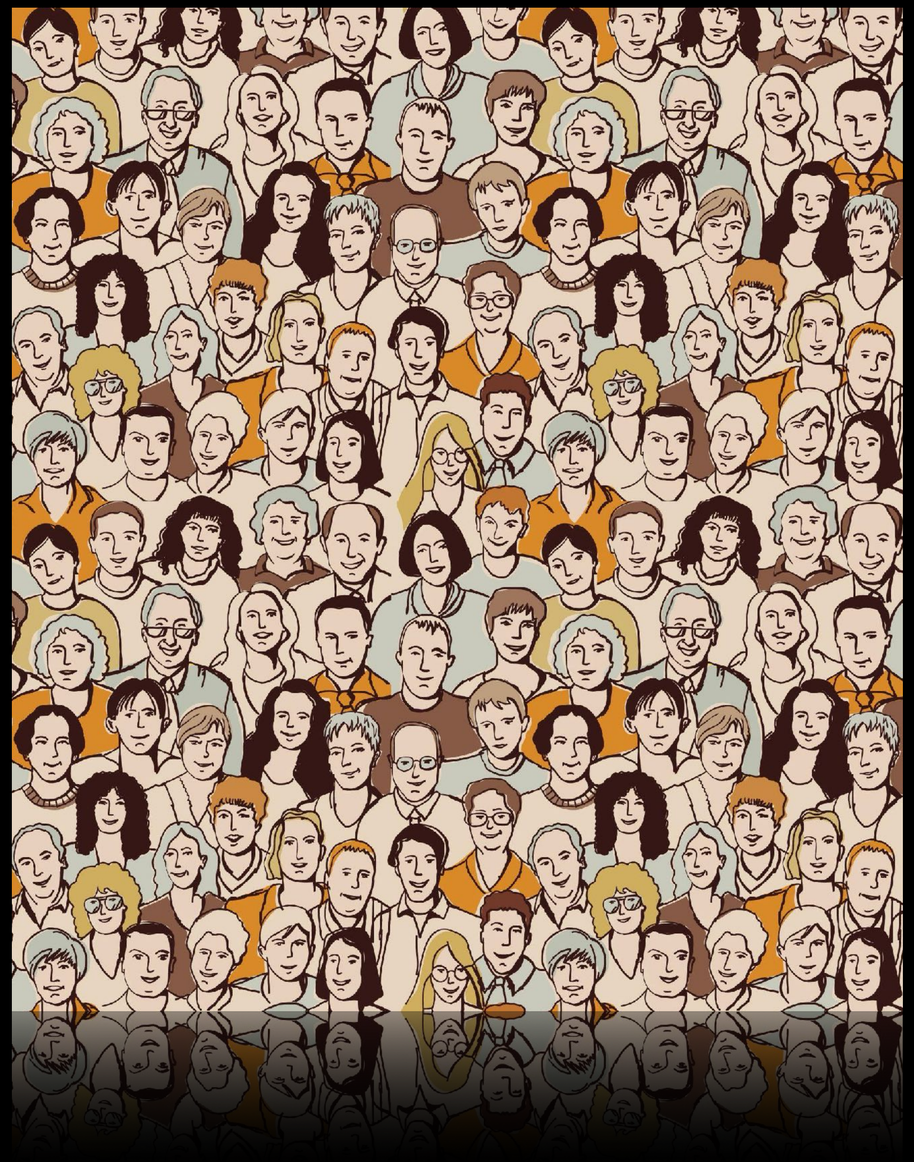

JL:

joon$ John Lin -> whisper
-> run for local mayor



I need to discuss this with my family, Eddy and Mei, and seek their input and support before I take any further steps in my political journey.

Generative Agent Architecture



Large language models can be prompted to generate human behavior conditioned on a variety of experiences.

GPT

`"[name] is a [description]"`

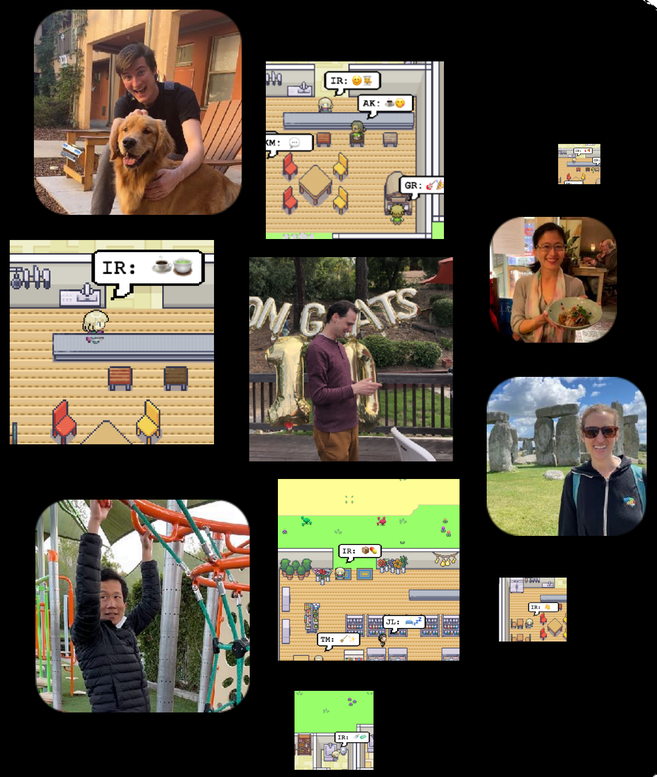
Social Simulacra (UIST '22)

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. UIST 2022.

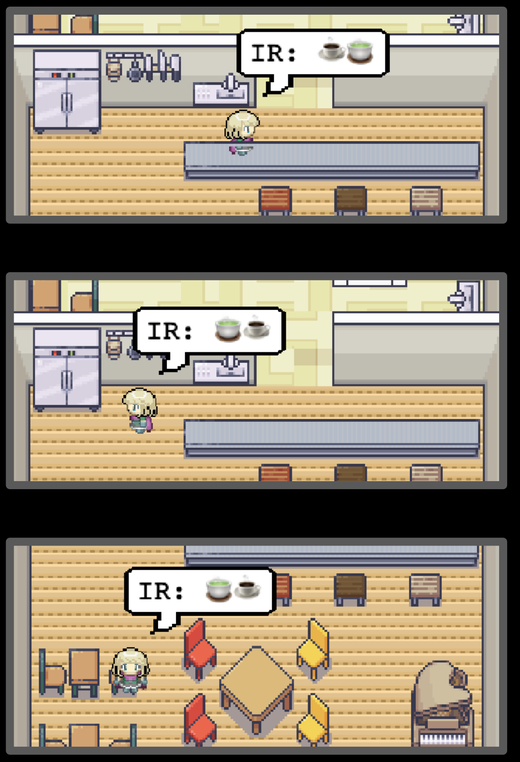
We remember and make sense of our experiences.

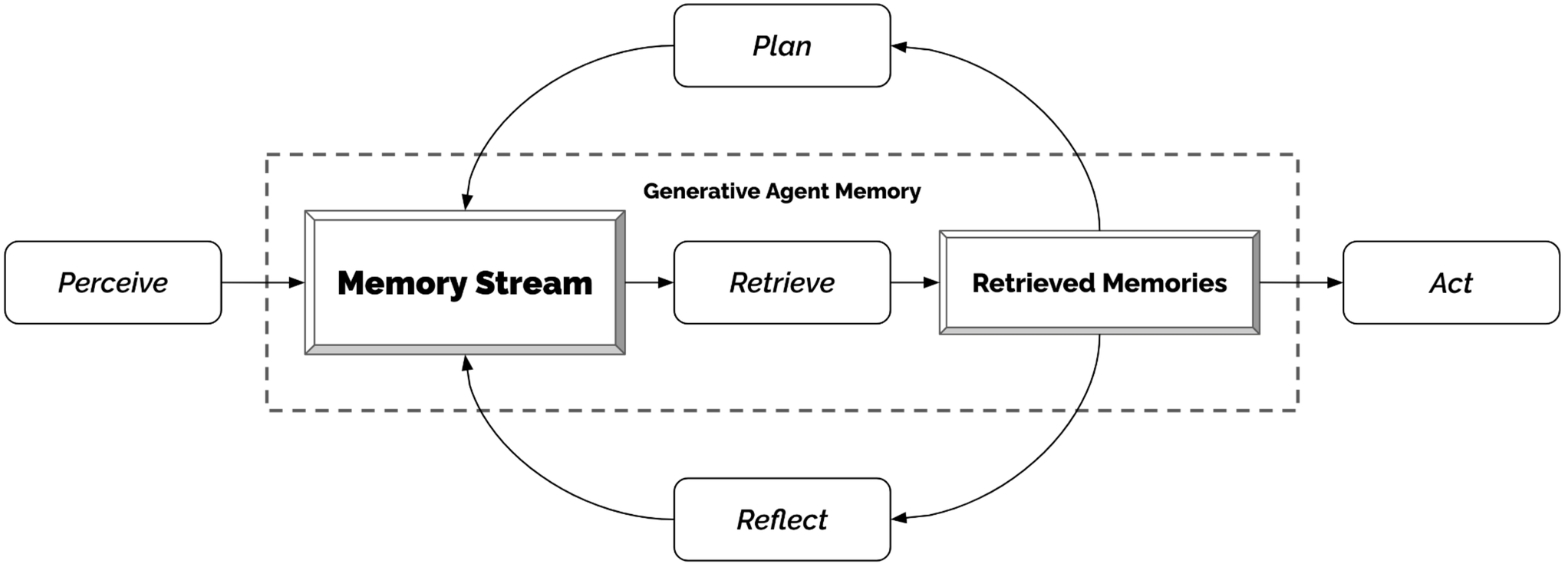
Prompt-based agents alone cannot.

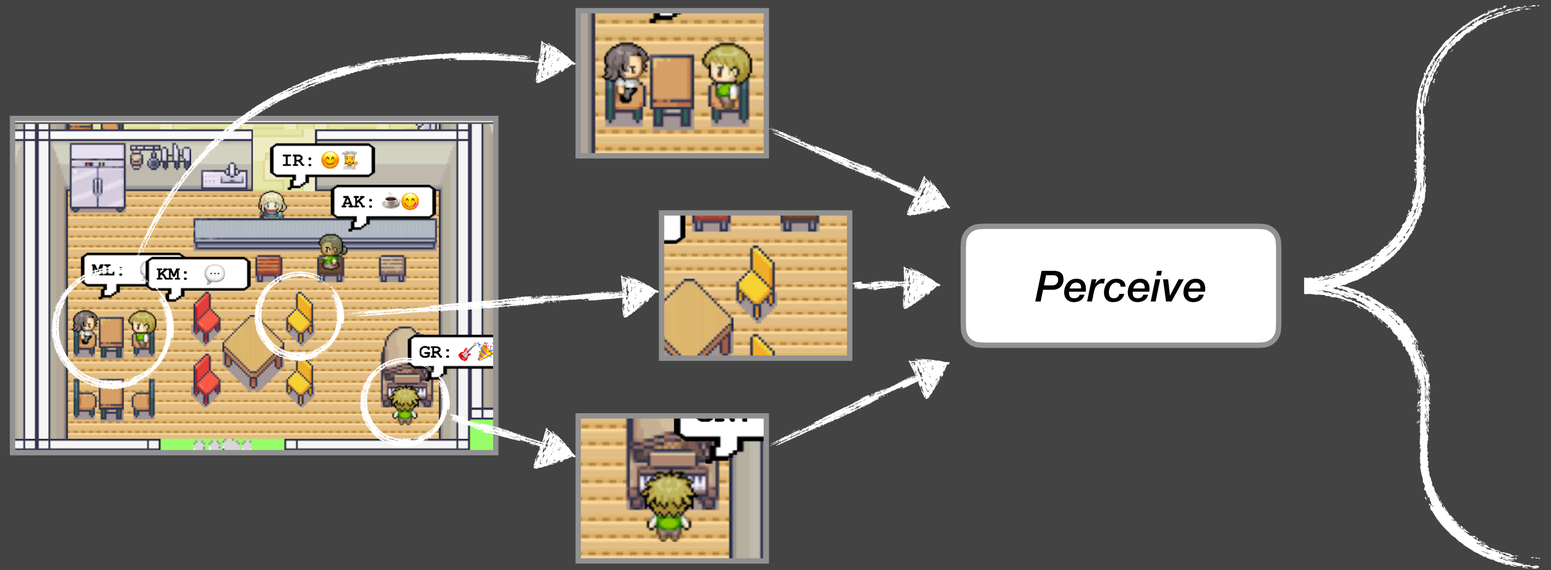
Perception

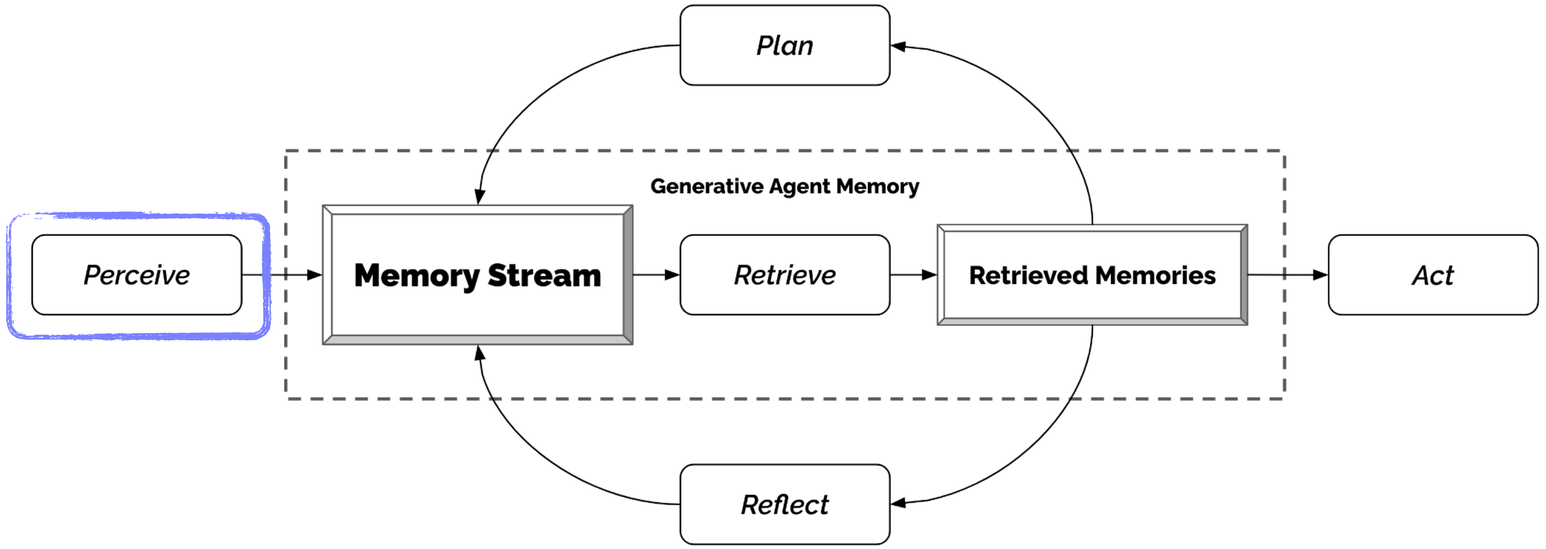


Action




Plan
Generative Agent Memory
Perceive
Memory Stream
Retrieve
Retrieved Memories
Act
Reflect

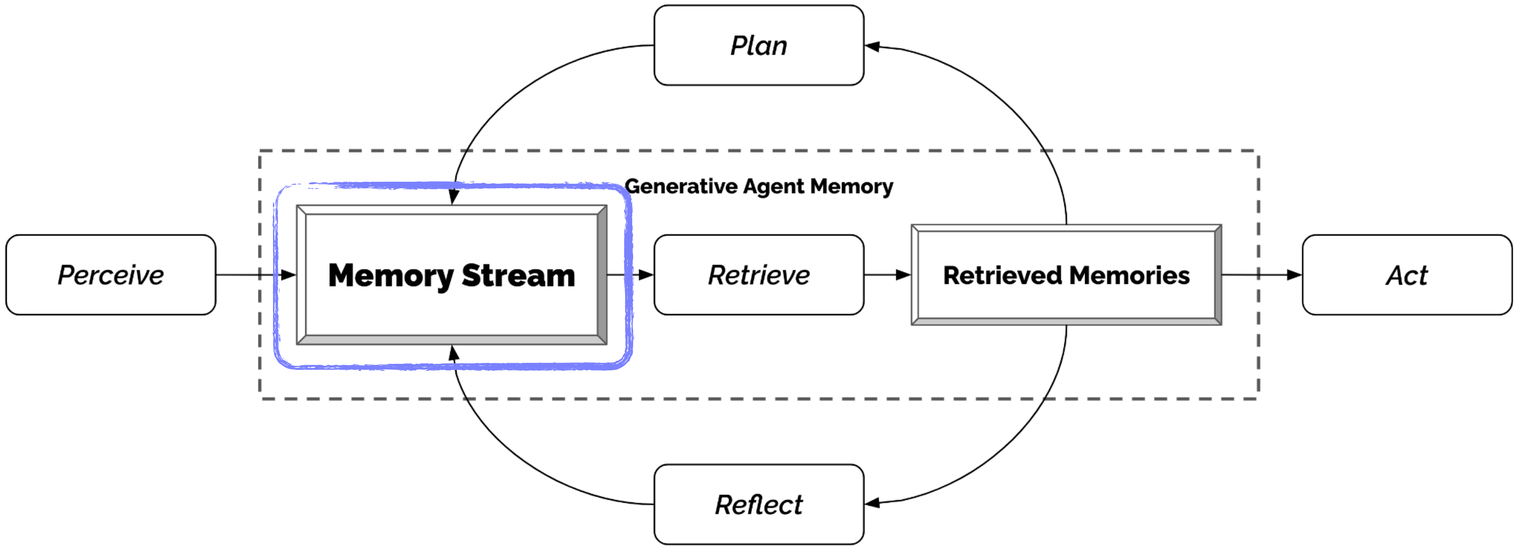

IR:
AK:
ML:
KM:
GR:
Perceive
Maria is chatting with Klaus
The chair is empty
Giorgio is playing the piano

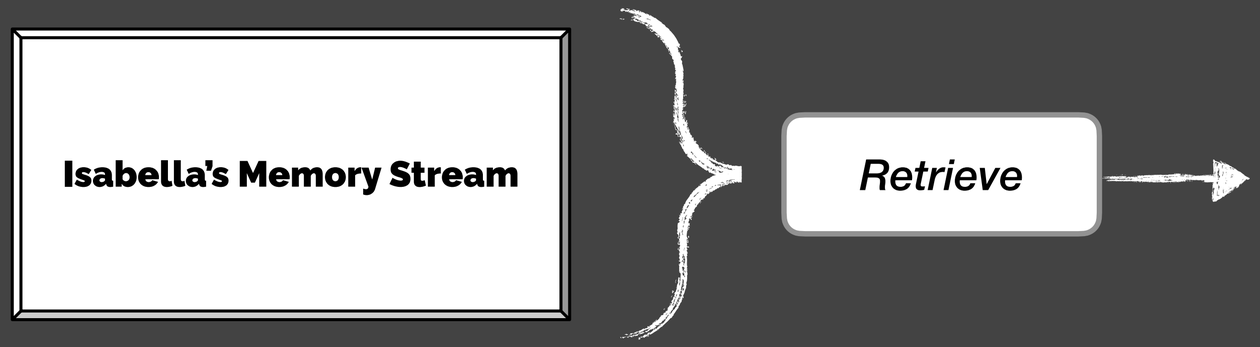

Plan
Generative Agent Memory
Perceive
Memory Stream
Retrieve
Retrieved Memories
Act
Reflect

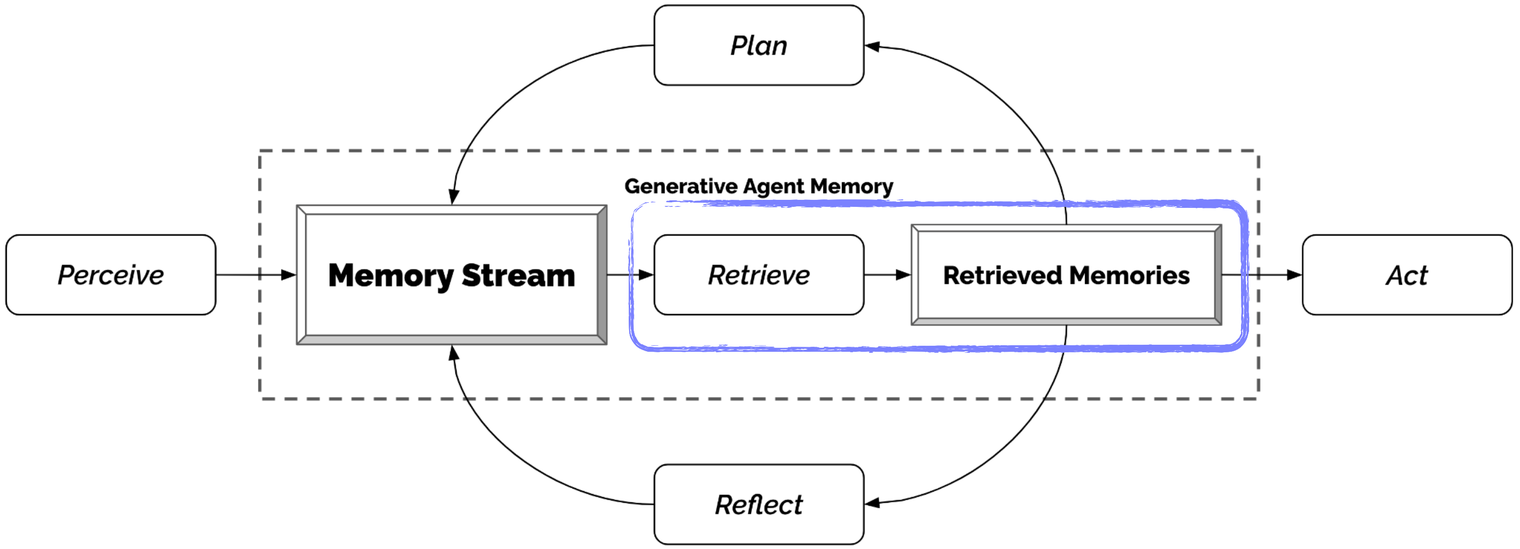
- Maria is chatting with Klaus
- The chair is empty
- Giorgio is playing the piano

Isabella's Memory Stream

2023-02-13 22:48:20: **Maria is chatting with Klaus**
2023-02-13 22:48:20: **The chair is empty**
2023-02-13 22:48:20: **Giorgio is playing the piano**
2023-02-13 22:48:20: Giorgio is playing the piano
2023-02-13 22:48:20: Giorgio is playing the piano
...




Isabella's Memory Stream
Retrieve
What are you excited about, Isabella?
Isabella is planning a Valentine's Day party at Hobbs Cafe.
ordering decorations for the party
researching ideas for the party


Plan
Generative Agent Memory
Perceive
Memory Stream
Retrieve
Retrieved Memories
Act
Reflect

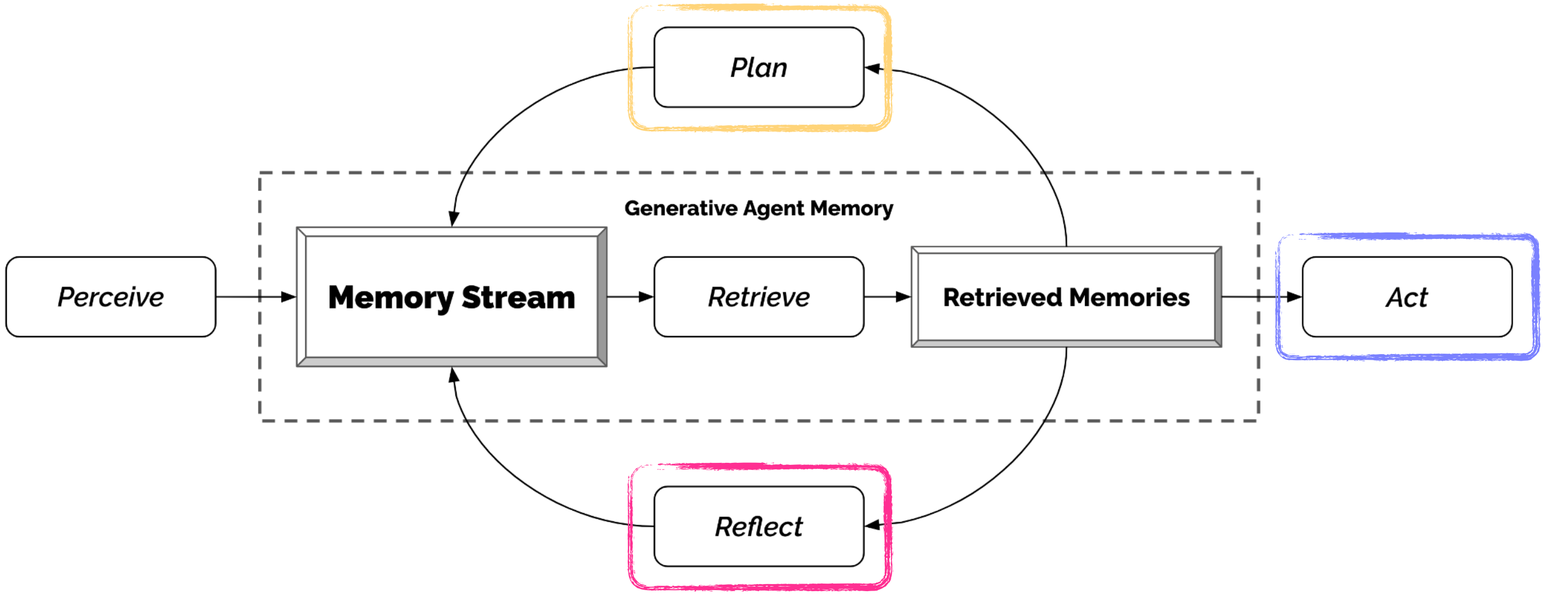
What are you excited about, Isabella?

- Isabella is planning a Valentine's Day party at Hobbs Cafe.
- ordering decorations for the party
- researching ideas for the party

[Plan] Let's decorate the cafe later this afternoon

[Action] Heading to the local grocery store to buy supplies for the party

[Reflection] I enjoy organizing events and making people feel welcome



To those in the cognitive architecture communities, these new architectures are immediately recognized.

Let me start by saying how much I enjoyed the Generative Agents paper. Really inspiring and imaginative work on how to leverage a LLM within social, interactive agents. There are many fascinating innovations in that work (for me), but what really got me thinking was the reflection section and how to use the LLM as an oracle for reflective knowledge and more broadly how all internal representations of knowledge can be in NL. Bravo to you and your students.

Cheers,

John Laird
(Professor Emeritus University of Michigan)

Simulation agents vs. tool-based agents

History repeats itself.

Early observation: scholars in cognitive psychology began to propose that the computers processed information similarly to human mind.

— Can we understand how human mind works by illustrating it with cognitive architectures?

— Can we create general-purpose computational agents that solve human tasks?

An interesting parallel:

Early observation: scholars in psychology and AI began to propose that the computers processed information similarly to human mind.

Classic cognitive architectures

Early observation: scholars in HCI and AI began to propose that generative AI encodes and generates human-like behaviors.

Today

But also...

— Can we understand how human mind works by illustrating it with cognitive architectures?

— Can we create general-purpose computational agents that solve human tasks?

Classic cognitive architectures

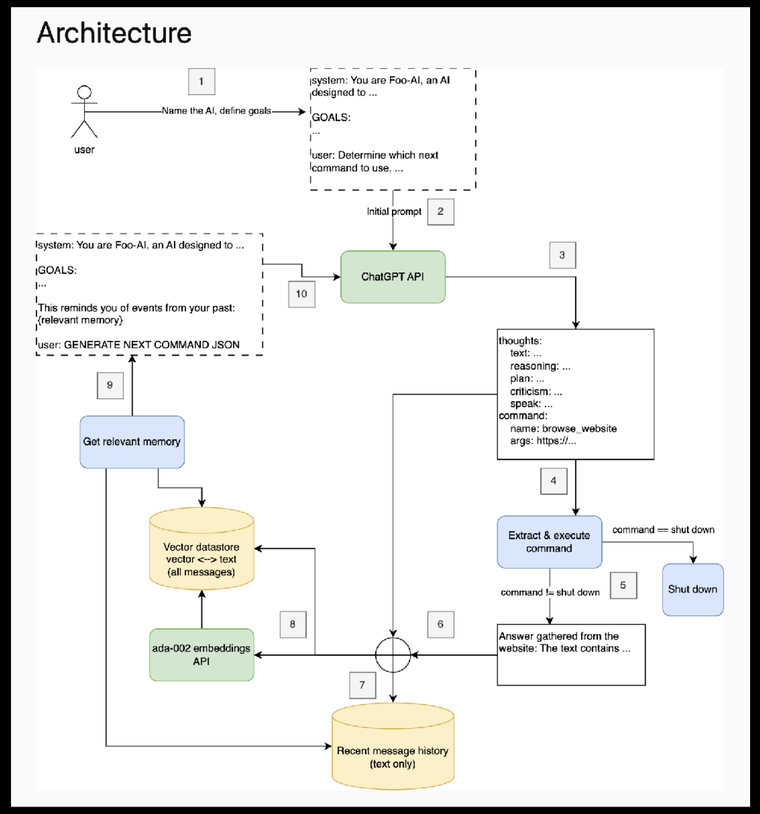
— Can we understand how people form emergent behaviors by illustrating it with generative agents?

— Can we create general-purpose computational agents that solve human tasks?

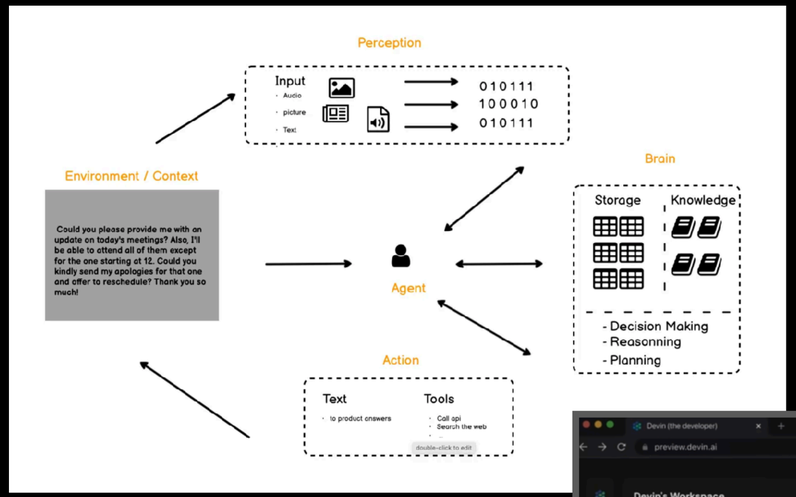
Today

Simulation agents vs tool-based agents

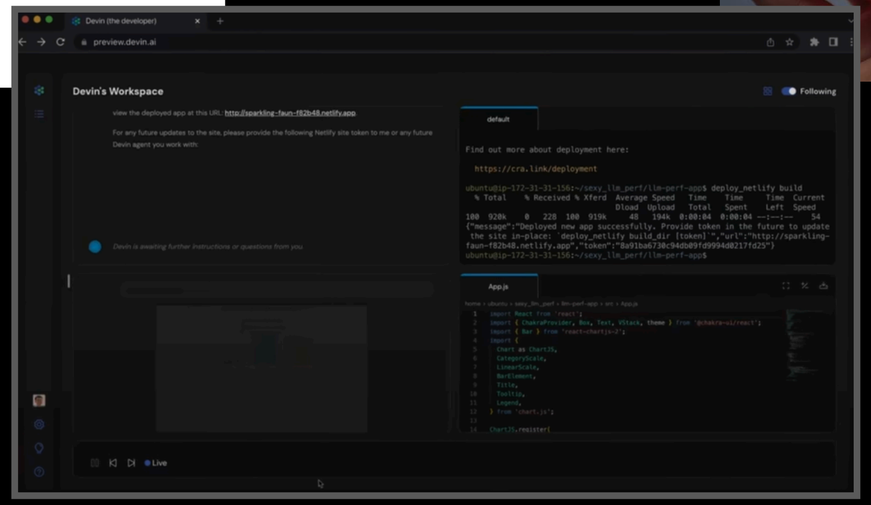
We are seeing different iterations of these architectures emerging in tool-based agents.



AutoGPT



Rabbit



Devin

Q. What are “architectures"?

https://pollev.com/helenav330

References

- M. Mathew, Understanding Operating System Architecture: Key Components and Features. Hashnode (2023); https://merwin.hashnode.dev/understanding-operating-system-architecture-key-components-and-features
- R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction (MIT Press, ed. 2, 2018).
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998-6008 (2017).
- J. F. Lehman, et al., A Gentle Introduction to Soar, an Architecture for Human Cognition: 2006 Update.
- SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).
- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023)
- A. Newell, Desires and Diversions (Carnegie Mellon University, Pittsburgh, PA, 1991).

References

- A. Newell, H. A. Simon, C. Shaw, The Logic Theory Machine. IRE Trans. Inf. Theory 2, 61-79 (1956).
- A. Newell, J. C. Shaw, H. A. Simon, Report on a general problem-solving program. Proceedings of the International Conference on Information Processing, 256-264 (UNESCO House, Paris, 1959).
- A. Newell, Unified Theories of Cognition (Harvard University Press, Cambridge, MA, 1990).
- J. E. Laird, A. Newell, P. S. Rosenbloom, SOAR: An architecture for general intelligence. Artif. Intell. 33, 1-64 (1987).
- J. R. Anderson, C. Lebiere, The Atomic Components of Thought (Lawrence Erlbaum Associates, Mahwah, NJ, 1998).
- Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. UIST 2022.

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park