Cognitive Architectures

LMARL Session 2, February 18th 2025, Polina Tsvilodub
In part based on slides of Michael Franke (SS 2018)

Recap

What are agents?

- agents combine many complex properties like the ability to achieve goals, autonomously solve novel problems, interact with an environment based on available information, interact with other agents
- goal-directed behavior can be decomposed into different sub-components
  - beliefs, desires, intentions / (sub)goals
- engineering approaches have strived to develop agents for different tasks, but flexible general systems have remained elusive
- cognitive science has provided insights into how humans navigate the complexities of real world

But how do we **build** systems with these features & capabilities?

What are architectures?

ARCHITECTURES

Cognitive architecture

> "[A] particular methodology for building [agents]. It specifies how . . . the agent can be decomposed into the construction of a set of component modules and how these modules should be made to interact. The total set of modules and their interactions has to provide an answer to the question of how the sensor data and the current internal state of the agent determine the actions . . . and future internal state of the agent. An architecture encompasses techniques and algorithms that support this methodology" (Maes, 1991, p.115)

Kaelbling considers an agent architecture to be:

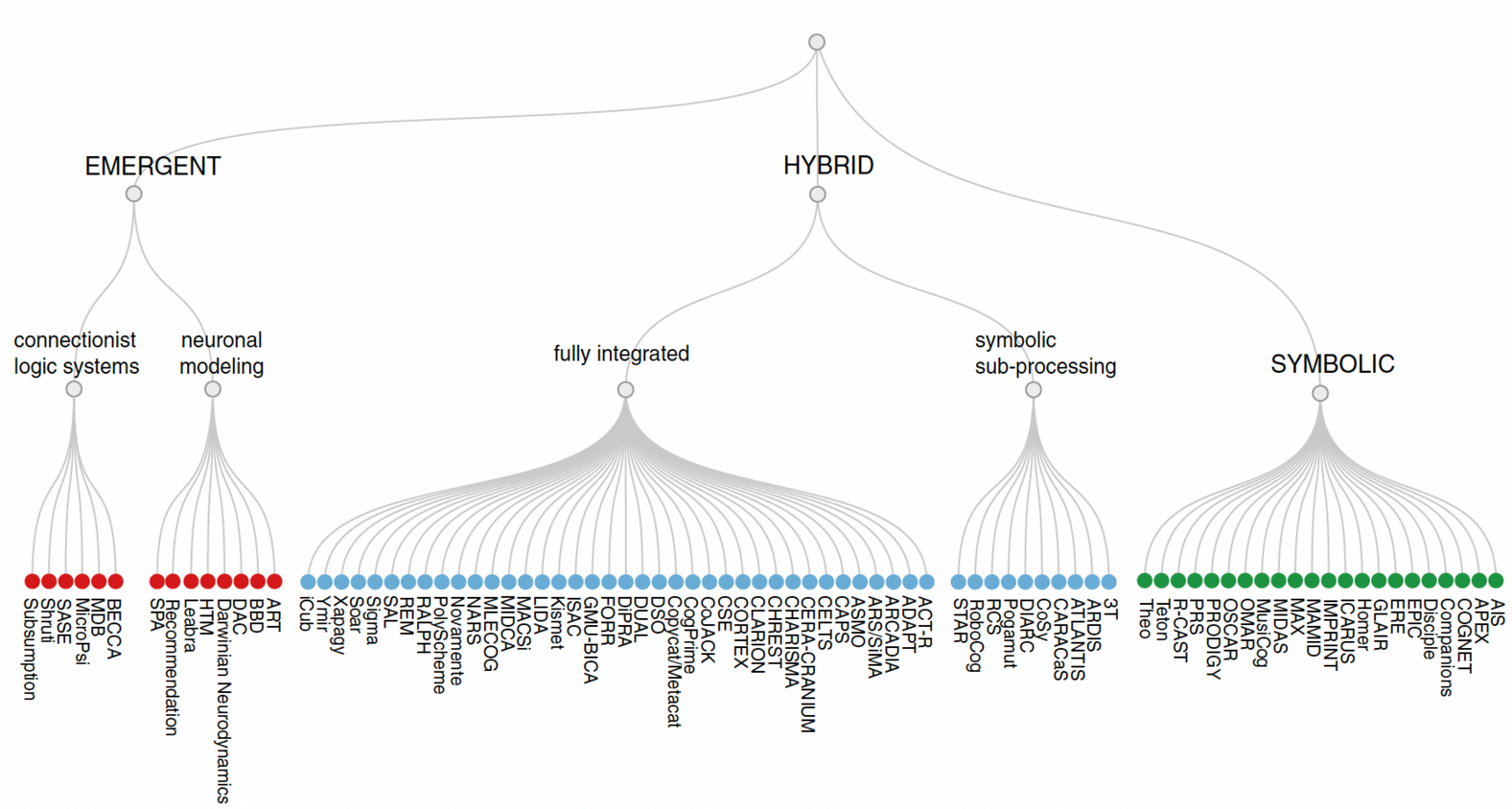
> "[A] specific collection of software (or hardware) modules, typically designated by boxes with arrows indicating the data and control flow among the modules. A more abstract view of an architecture is as a general methodology for designing particular modular decompositions for particular tasks." (Kaelbling, 1991, p.86)

cognitive architectures are both a functional description and a theory of humanlike minds, designed to model such minds and their functionality leading to intelligence

Wooldridge & Jennings (1995), Laird et al. (2017)

Cognitive architecture

- cognitive architectures are both a functional description and a theory of humanlike minds, designed to model such minds and their functionality leading to intelligence
- assumption of a **computational theory of mind**



Kotseruba & Tsotsos (2018) Laird et al. (2017)

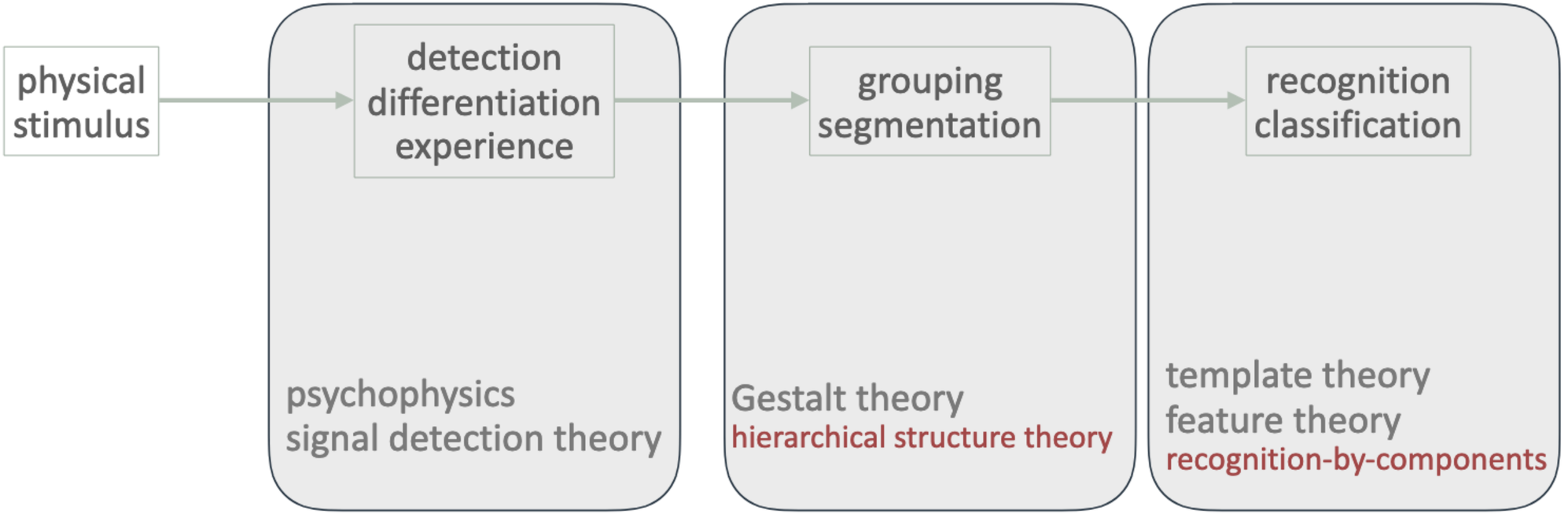
Cognitive architecture

Overview of general components:

- **Perception**: transform raw sensory input into representations
- **Attention Mechanisms**: allocate cognitive resources to certain information
- **Action Selection**: decision-making processes as to which actions to undertake
- **Memory Systems**:
  - Short-term memory
  - Working Memory: "temporary storage" for active tasks
  - Long-Term Memory: “permanent” storage for knowledge, rules, and experiences
    - episodic memory
    - procedural memory
    - semantic memory
- **Learning**:
  - adaptation, generalization based on experiences
- **Reasoning & Metacognition**:
  - Logical inference, probabilistic models, DPT, self-reflective thinking

Laird et al. (2017)

Perception



- forming an internal representation of a co-present physical stimulus
- perception is an active process influenced
- by context and knowledge: top-down vs bottom-up processes
- perception is similar to decision making: think of perceptual decisions
- though stimuli are often fuzzy and vague, perception is often stable and categorical
- role of features or components for noise-robust rapid recognition
- perception is heavily influenced by context, expectation and interference from other sensory modalities

Attention

::: selective allocation of limited processing resources (at the cost of processing other stimuli)

- what decides where attention goes?
- where is the limit of equal parallel processing?
  - i.e., when does attention have to select what to focus on?
- what happens to what is outside of our attention?

There are different accounts of attention:

- stimulus-driven vs. goal-driven attention
- early vs. late filter theories
- attenuation theory, neglect

Human memory

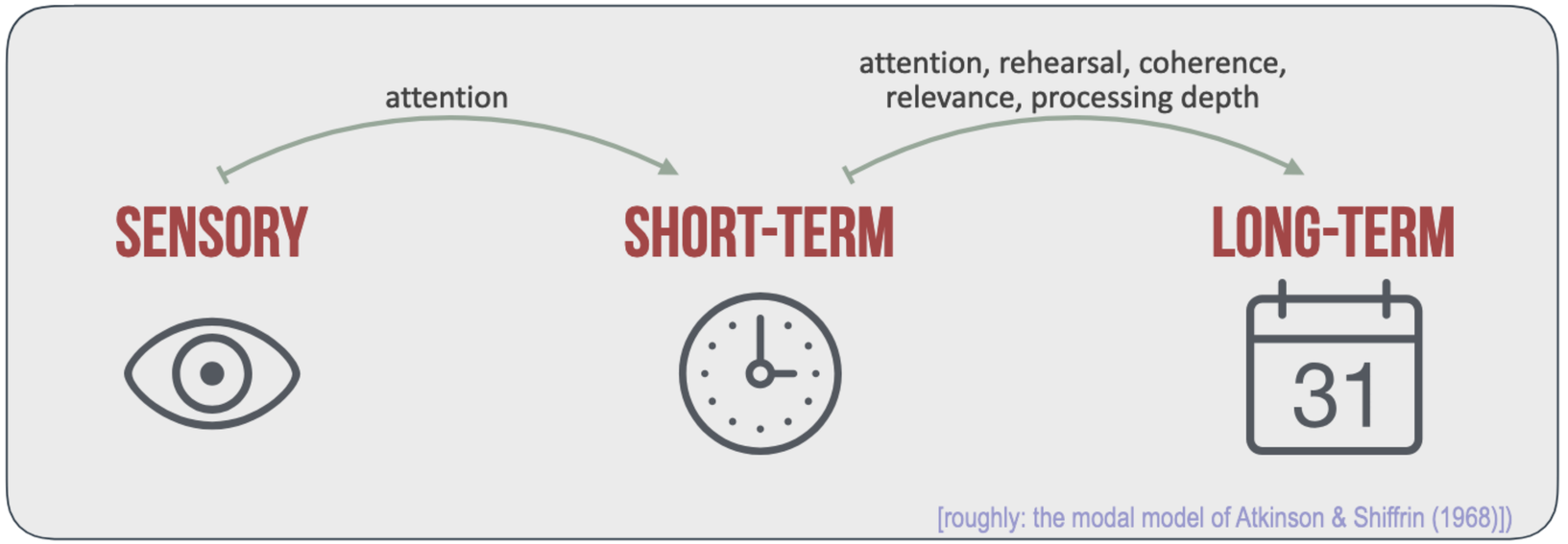


TABLE 5.2 Agreement About the Actions Stereotypically Involved in Going to a Restaurant

Open door[a]
Enter[b]
Give reservation name
Wait to be seated
Go to table

Sit down[c]
Order drinks
Put napkins on lap

Look at menu
Discuss menu

Order meal
Talk
Drink water
Eat salad or soup
Meal arrives

Eat food
Finish meal
Order dessert
Eat dessert
Ask for bill
Bill arrives

Pay bill
Leave tip
Get coats

Leave

[a]Roman type indicates items listed by at least 25% of the participants.
[b]Italic type indicates items listed by at least 48% of the participants.
[c]Boldface type indicates items listed by at least 73% of the participants.

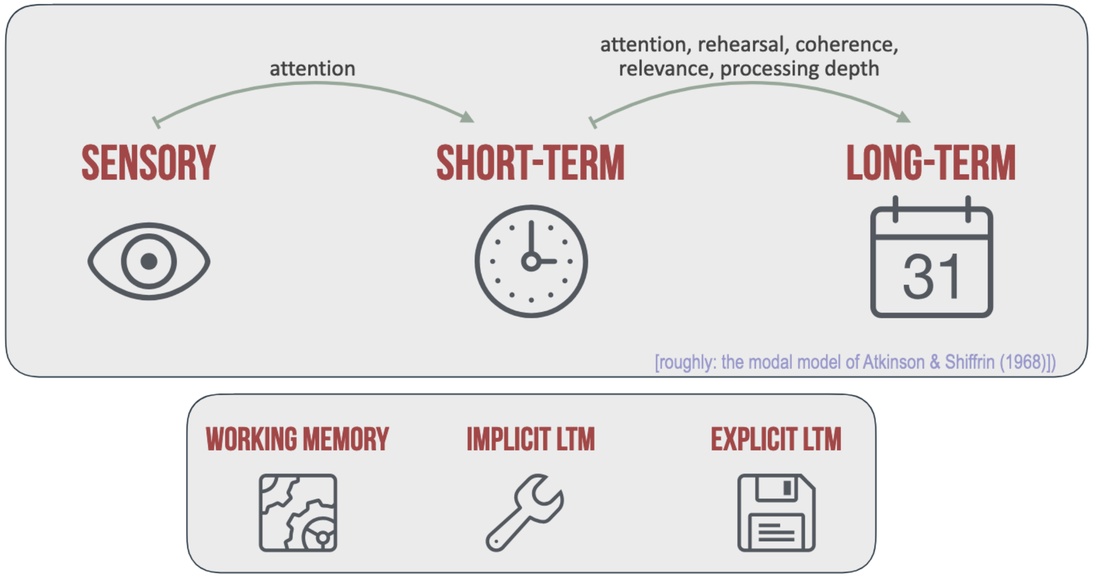
Adapted from Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177–220. Copyright © 1979 Elsevier. Reprinted by permission.

Anderson (textbook), pages 116—118

Schank & Abelson (1977)

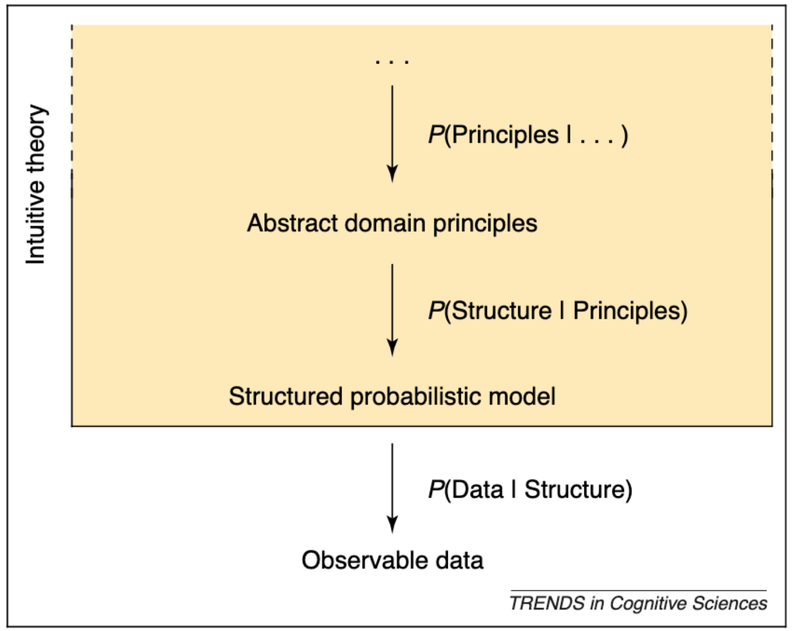
Learning
Based on memory

- associative vs. non-associative
- explicit and implicit
- transfer learning
- social learning



Bayesian models of inductive learning

$$P(h|x,T) = \frac{P(x|h,T)P(h|T)}{\sum_{h' \in H_T} P(x|h',T)P(h'|T)}$$

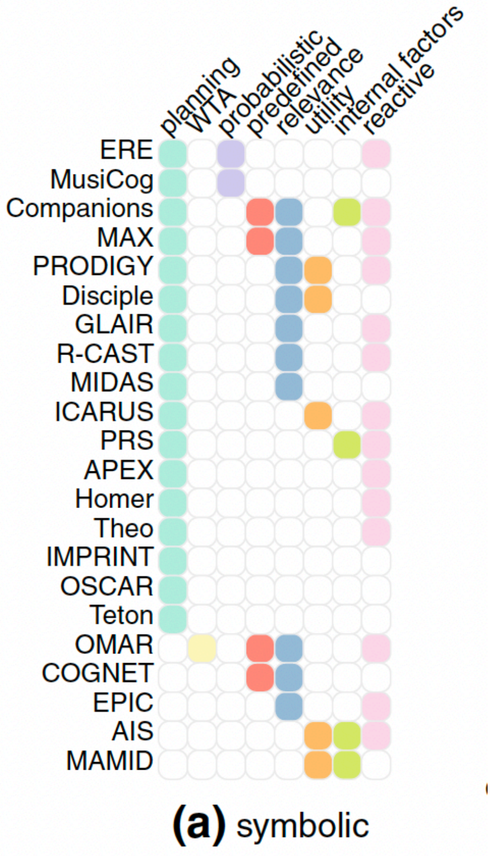


Tenenbaum et al. (2006)

Action selection

Policy

- **planning**: given G, decide on sequence of actions that will lead to passing goal test, e.g. by recursively breaking down the goal
- **reactive action selection**: suspend current activity, immediately react to stimulus
- dynamic action selection:
  - Winner-Take-All
  - probabilistic
  - predefined order
- criteria
  - relevance
  - utility
  - internal factors



(a) symbolic

Kotseruba & Tsotsos (2018)

Reasoning

- explicit (logical), probabilistic
- dual process theory
- practical vs. epistemic

PURE/INSTRUMENTAL

- derived from logic, probability theory, decision (game) theory etc.
- provides a normative standard for how people should behave in the abstract
- demands that people abide by this standard in every laboratory situation (no matter whether natural/familiar/etc. or not)

ECOLOGICAL/ADAPTIVE

- speculate about general purpose/function of some cognitive function/behavior
- consider the environment to which cognition is adapted (≠ lab experiments)
- rationalize experimental results as reflex of a general pattern of behavior that is a good adaptation (possibly heuristic) elsewhere

Citation (2002), Citation 2 (2050)

General Problem Solver

- “general intelligence” system
  - 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
  - assumption: a physical symbol system has the necessary and sufficient means for general intelligent action
- core idea: use means-ends analysis to reduce difference between current and goal state

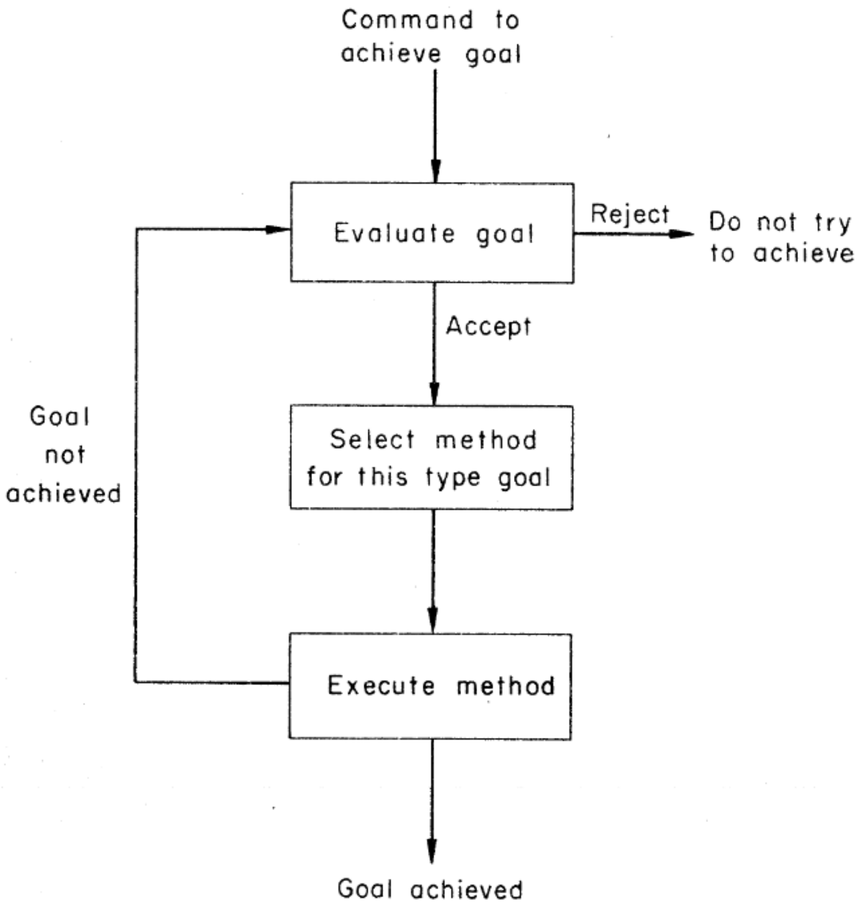


Fig. I — Executive organization of GPS

Newell, Shaw & Simon (1959)

General Problem Solver

Means-ends analysis

- define problem (current and goal state), **specification of pre- and post-conditions of actions**
- compare current and goal states
- identify differences
- select an operator to reduce the difference
- apply the operator and repeat until goal is reached

What are the assumptions?

- structured state and action representations
- respecified set of goal types and operators

- **Problem solving as a search over structured problem space**

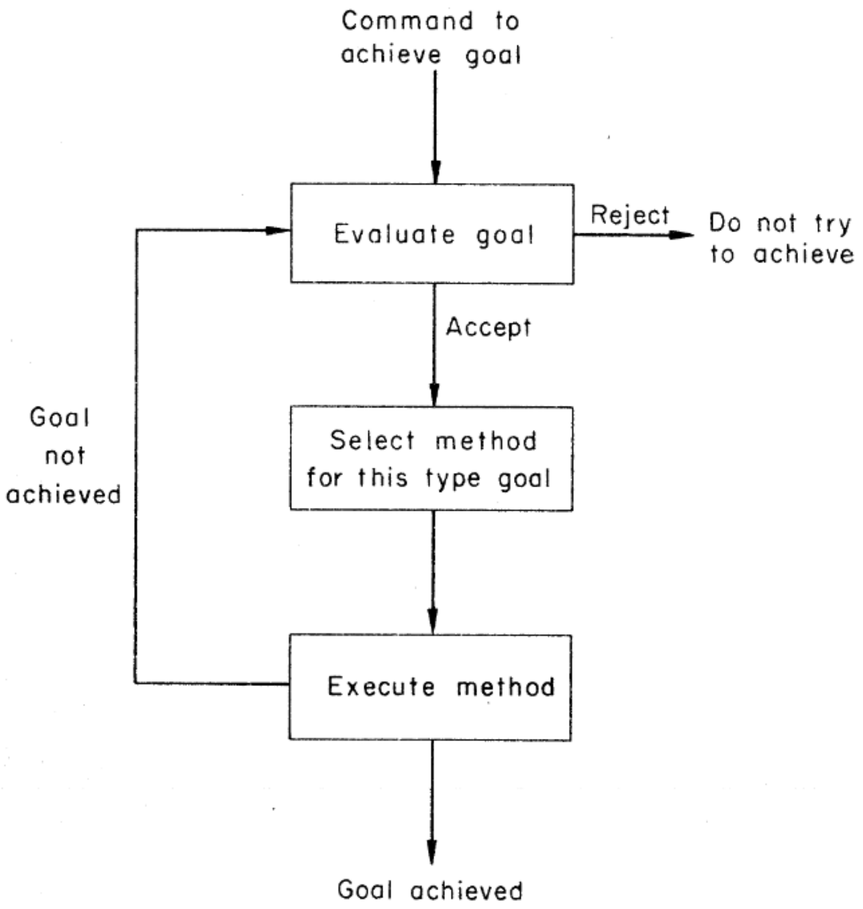


Fig. I — Executive organization of GPS

Newell, Shaw & Simon (1959)

General Problem Solver

Means-ends analysis

Strengths	Limitations
A foundational step towards automated problem-solving	Computational inefficiency
Generalizable across multiple domains	Struggled with complex, real-world problems
Inspired modern AI techniques, including planning and reasoning	Required structured formalized problem definitions

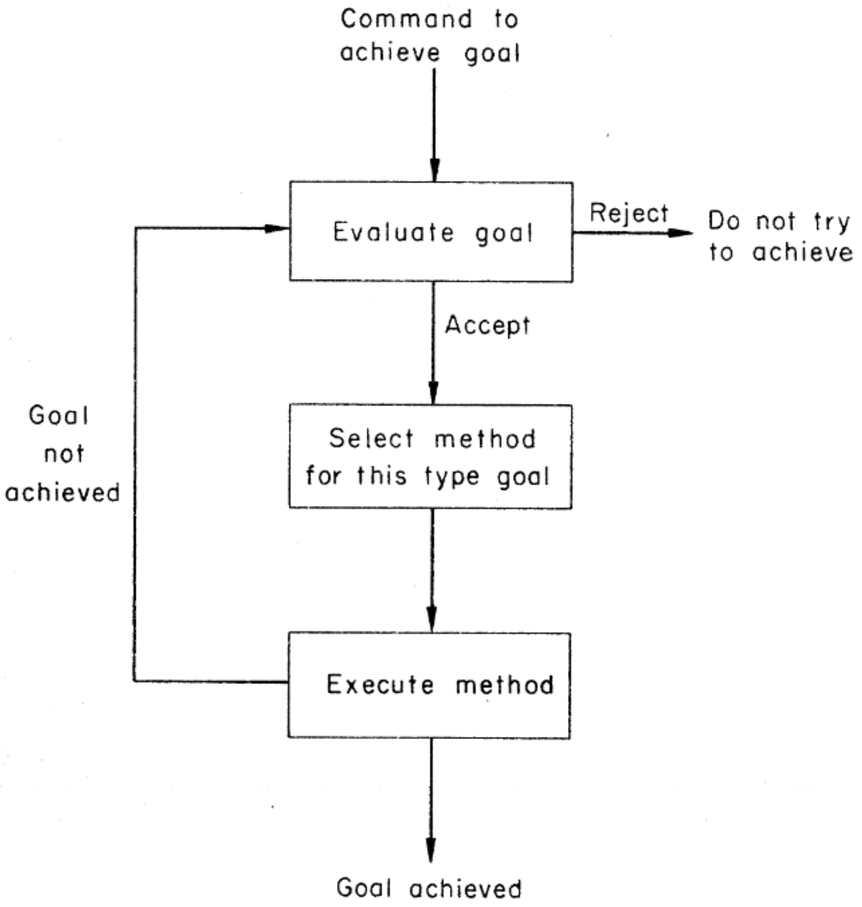


Fig. I — Executive organization of GPS

Newell, Shaw & Simon (1959)

Generative Agents as Simulacra of Human Behavior

LLMs as knowledge bases (for agent simulations)

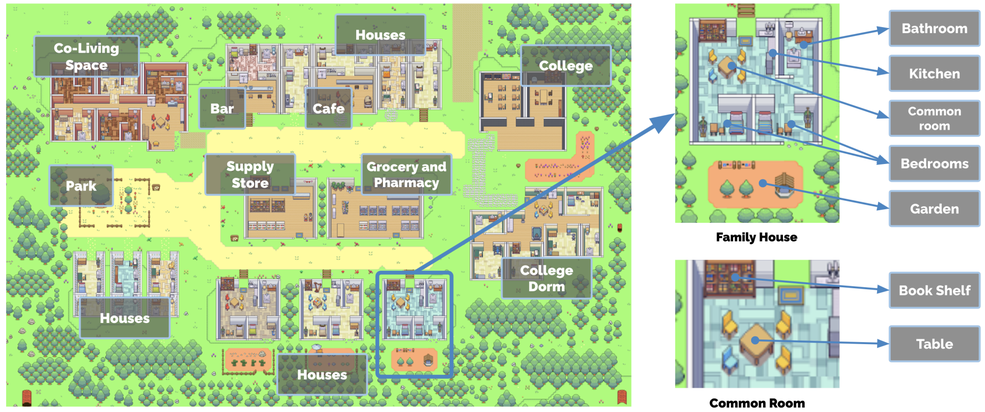
"The key observation is that large language models encode a wide range of human behavior represented in their training data. […]."

"[…] we compare GPT-4 to ChatGPT throughout to showcase a giant leap in level of common sense learned by GPT-4 compared to its predecessor."

Park et al. (2023), Microsoft (2023)

Generative agents
as simulacra

- The Sims-style environment Smallville in which LLM based agents dynamically simulate human behavior
- motivation: realistic non-human players in games
  - LLM-based agents dynamically interact w/ each other
- based on 25 agents (initialized with text bio)
  - interaction with environment via descriptions of actions
  - (emergent) social behavior between agents
  - user intervention via conversation or direct instruction
  - game sandbox movements computed based on LLM output

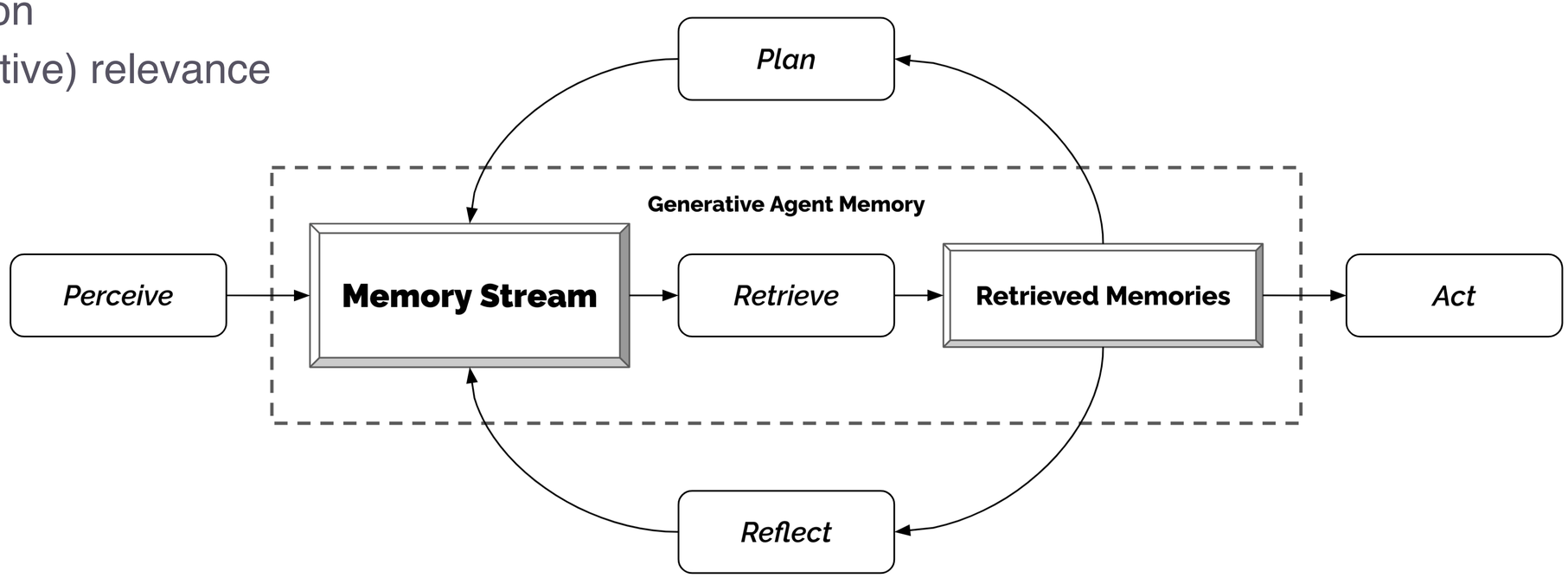


Park et al. (2023)

Generative agents

player model

- complex agent model
  - informed by human psychology
    - planning
    - memory (long / short)
    - reflection
    - (subjective) relevance



Park et al. (2023)

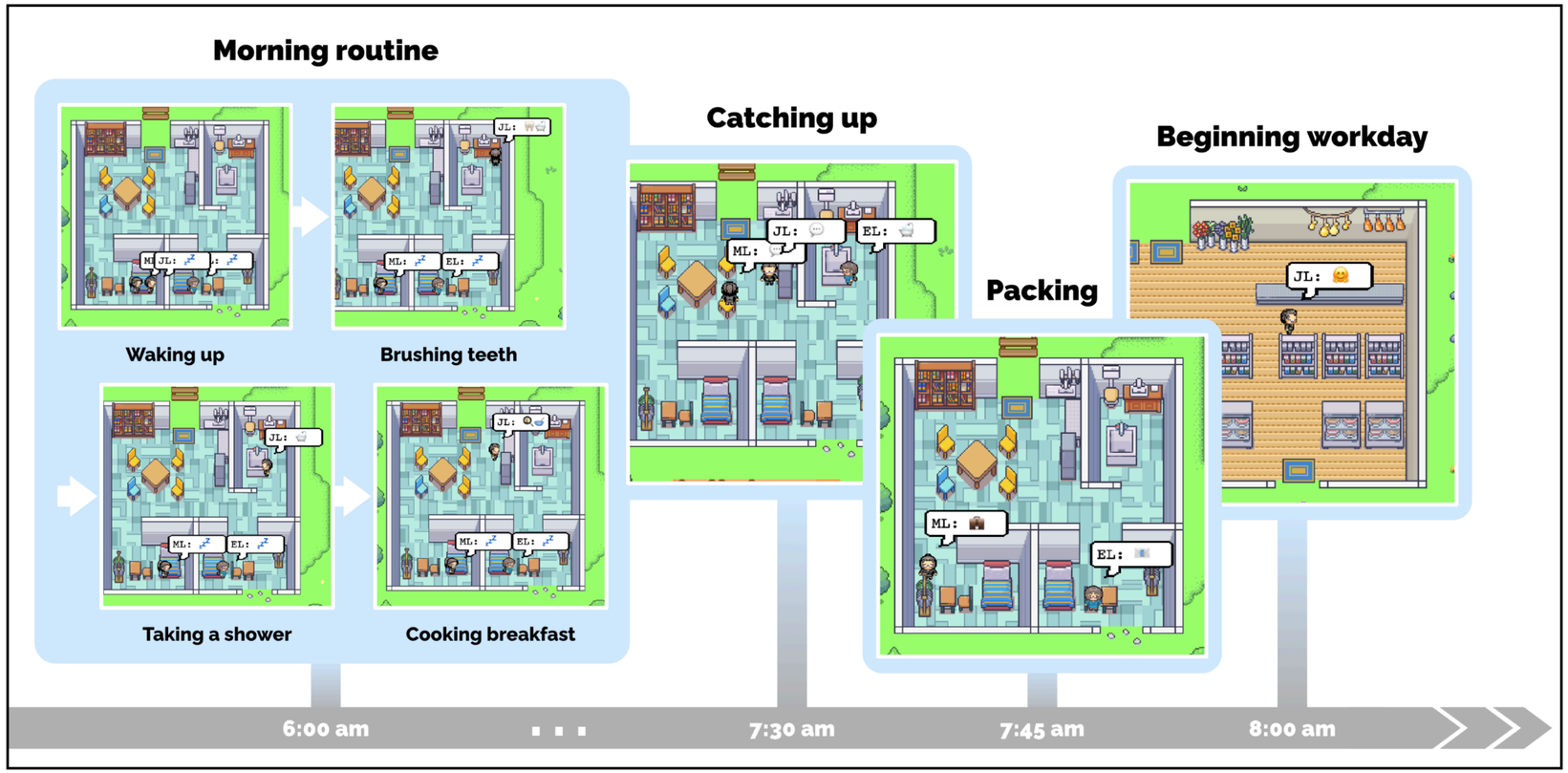
Generative agents

player characters and interactions

agent character descriptions

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well — the husband Tom Moreno and the wife Jane Moreno.

interaction w/ environment & other characters

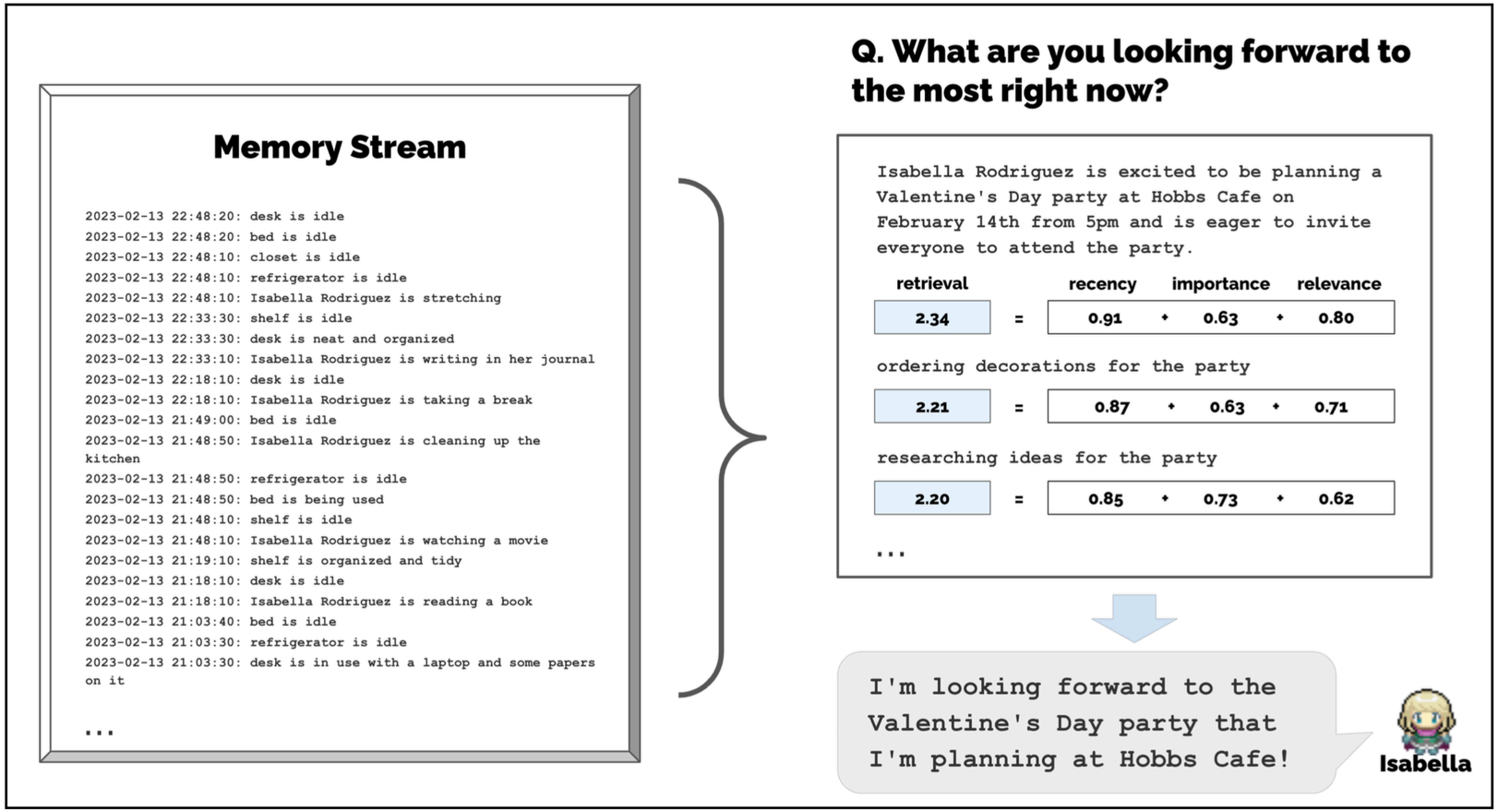


Park et al. (2023)

Generative agents

player model ::: relevance-based retrieval

filtering raw memories based on relevance



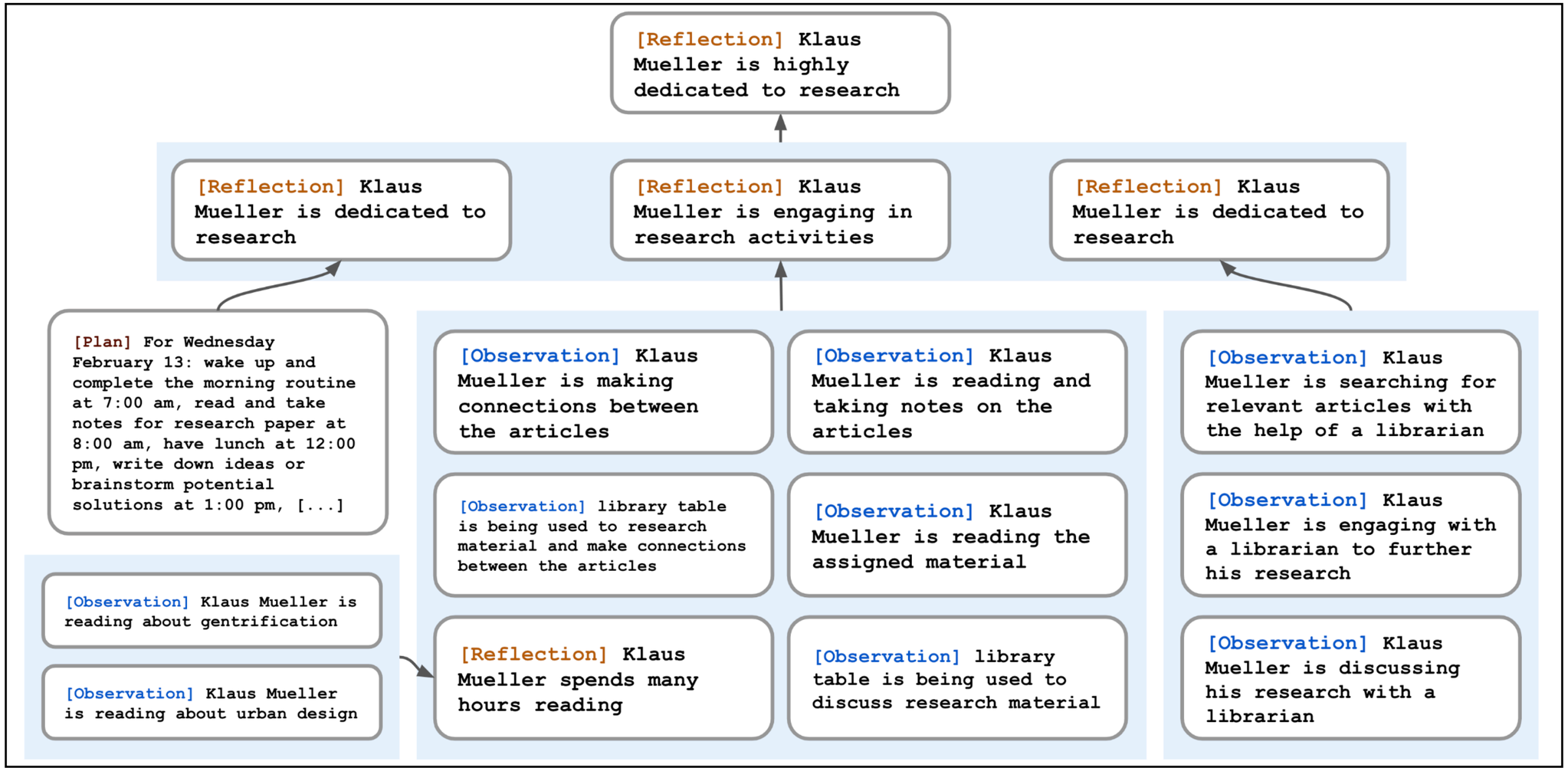
LLM-prompt relevance judgement

On the scale of 1 to 10, where 1 is purely mundane (e.g., brushing teeth, making bed) and 10 is extremely poignant (e.g., a break up, college acceptance), rate the likely poignancy of the following piece of memory.
Memory: buying groceries at The Willows Market and Pharmacy
Rating: <fill in>

Park et al. (2023)

Generative agents

player model ::: self-reflection



Park et al. (2023)

Paper discussion logistics

- **First discuss questions from Moodle!**
- What are the core modules that constitute a humanlike mind?
- What do you think is an optimal level of granularity of decomposition of the modules in the architecture and why?

1. split in two groups, half of the experts in each group
2. discuss the paper (e.g., start with questions of non-expert participants)
3. experts: responsible for adding key points, insights, new questions of the group to shared Google slides: https://docs.google.com/presentation/d/1BH53A2ipfzrix9C0gR39Cd1uUIHZaIZ2_CseZmhe81U/edit?usp=sharing
   a. maximally 3 slides!
   b. make slides such that they will be helpful for exam!
4. joint discussion