

AI in Context 2025, [valueslab.github.io](https://valueslab.github.io)

# AI and Memory

Katja Maria Vogt

ValuesLab

## Plan for Today

Memory is a major topic in empirical psychology and neuroscience, and an emerging topic in research on AI. Plan for today:

- What is called memory in AI?
- Why is this important? Benefits and dangers.
- Does the use of AI affect human memory?
- We'll discuss a range of concerns w.r.t. memory that were raised already in antiquity, and see how they relate to AI.

## Human Memory and AI

Talk about memory in AI comes up in uses of LLMs where the AI “remembers,” for example, earlier moves in:

- Conversations with AI companions, AI therapy, AI tutoring, etc.
- Robots with an LLM component, e.g. elder care robots.

## Human Memory (with thanks to Elliot Blake!)

- Memory is one of the default activities in the human brain.
- Memory is connected to a wide range of other cognitive activities, such as belief, planning, intention, imagination, etc.
- The opposite of remembering is forgetting.
- Forgetting is a failure, but not all forgetting is bad (trauma, etc.).
- Forgetting can occur at any stage of the three key memory processes: acquisition, retention, or retrieval.

## More on Human Memory (with thanks to Elliot Blake!)

- Sensory memory: momentarily retains raw sensory information for the brain to process.
- Working memory: retention of information for duration of a task.
- Long-term memory:
  - Non-declarative/implicit
  - Declarative/explicit: episodic, semantic, reference

## Episodic and Semantic Memory

- Episodic memory: of events that involve the agent herself.
- Semantic memory: of (other) facts about the world, including of past events that one did not experience oneself.
- Contested case: you recall something about your childhood because your parents told you about it. This seems semantic, not episodic, but it is about your own past.

## Memory—Following Michaelian and Sutton (2017)

Prospective memory:

“Prospective memory refers to the ability to remember to perform a planned action, or to execute an intention.”

Note: While memory seems to be past-directed, it can also be future-directed, namely, when one remembers a plan.

## Memory and Human Capabilities

Narrative theories of the self assume that we “tell ourselves a story” about who we are (Rosati 2013). These stories

- aim to get right what happened
- shift over time
- have foci depending on the occasion/time/interlocutor/etc

Upshot: Episodic memory isn’t simply information-storage. It involves imaginative activity, relates to self-conceptions, etc.

Question: If robots—“AI agents”—have memory-like capabilities, will they also engage in this kind of constructive work?



## Memory in AI

For LLMs used in therapy, elder care robots, etc., it matters greatly how much the LLM can retain from earlier conversations.

- Short Term Memory (STM): Duration of a conversation.
- Long Term Memory (LTM): Persistent across sessions.
  - Context window: roughly, how much information an LLM can handle at once.

## Memorization vs. Generalization—Following Tirumala (2022)

- AIs work by next token prediction. This involves:
  - *Generalization* over lots of data
  - *Probabilistic* prediction
- This is fundamentally different from memorization. Why?
  - Memorization is token by token: *this* sentence, *that* photo, etc.
  - Generalization & probabilistic prediction: *many* tokens.

## Memorization vs. Generalization—Following Tirumala (2022)

- The aim of generalization & probabilistic prediction is to generate novel outputs that concern tokens that were *not* in the training set.
- Overfitting: cases where the AI learns the details, including “noise,” of a training set so closely that it fails to generalize w.r.t. new items.

Example: you want the AI to recognize dogs. The AI is trained with a small set that is “noisy”; for example, all dogs on the photos in the training set sit/run in gardens. Then you show a photo of a dog in a living room to the AI. The AI doesn’t recognize the dog as a dog.

## Memory in AI—Following Tirumala 2022

- Unintended memorization: LLMs quote some of the training data.
- This is a *failure* to generalize.
- It happens that an LLM quotes an entire article that was online.
- This is most sensitive w.r.t. quoting personal information.
- This “makes it vulnerable to extraction attacks,” where training data are extracted from LLMs.

## Memory and Forgetting—Following Tirumala 2022

Forgetting in AIs?

“There has also been work studying memory degradation (forgetting) in language models. [...] neural networks tend to forget the information from previous trained tasks or training batches, when trained on new data.”

## Memory and Unlearning—Following Tirumala 2022

### Unlearning in AIs?

“Machine unlearning is a technique that forces a trained model to forget a previously learned sample [12, 54], which is primarily motivated by data protection and privacy regulations.” (p.2)

## Memory and AI

Does the use of AI affect our capacity for memory?

- This question concerns, in the first instance, semantic memory.
- The concern is that, over time, the use of AI negatively affects our ability to remember things.

Cognitive cost of using LLMs in essay writing:

- Three groups: LLM group, Search Engine group, Brain-only group.
- Brain-only group: highest memory recall, most wide-ranging and strongest brain activity.
- LLM group: fell behind on ability to quote from their essays.



Plato, *Phaedrus* 274c-276a

- Earlier versions of this concern are familiar, say, w.r.t. use of the internet.
- A much earlier version: the use of writing (sic!) comes with cognitive costs.
- Plato entertains this in the dialogue *Phaedrus*, where Socrates tells a story about the invention of writing in Egypt.
- The core concern is that writing will negatively affect memory.

## Invention of Writing in Greece

- Linear B, syllabic signs: 1400–1200 BCE
- Alphabetic writing: 8th century BCE
- Writing down of Homer's epics c. 8th century BCE

## Homer Before Writing?

- How did people compose and remember complex epic poems that, in today's editions, run over several hundreds of pages?
- Mnemonic techniques: practices designed to enhance memory.
- Compositional techniques: for example, ring composition, designed to structure the progression of the story in ways that aid memory.

## Homer Before Writing?

- Did people have better memory? Hard to say!
- The most famous text, in this regard, is the so-called Catalogue of Ships in Homer's *Iliad* 2.484-759.
- List of the Greek troops sailing to Troy, ship by ship, described by who the leader is, etc.
- This is so famous that the Hollywood movie *Troy*, otherwise more interested in romance and fighting, devotes pride of place to it.
- <<https://www.youtube.com/watch?v=cFuXesZfbnA>>

## Concerns in Plato's *Phaedrus*—AI Parallels?

- The (presumed) inventor of writing, Theuth, claims that writing will improve people's memory.
- Thamus, the Egyptian king, says that Theuth, as inventor of writing, has affection for it and misjudges its effects.
- Those who invent X and those who can assess X are not the same.
  - AI-1: Is this a valid concern? Are those who develop AI, qua inventors, too much in love with AI to assess it?

## Concerns in Plato's *Phaedrus*—AI Parallels?

Thamus: “it will introduce forgetfulness into the soul of those who learn it: they will not practice using their memory because they will put their trust in writing” (275a)

- AI-2: Is there a risk that one becomes forgetful if one
  - doesn't *practice memory* and
  - puts one's *trust* into the written replies one receives?

## Concerns in Plato's *Phaedrus*—AI Parallels?

Thamus: “Your invention will enable them to *hear many things without being properly taught*, and they will imagine that they have come to know much while for the most part they will know nothing. And they will be *difficult to get along with*, since they will merely appear to be wise instead of really being so.” (275a-b)

- AI-3: What is the difference between receiving an answer to a prompt and being properly taught?
- AI-4: Will people with access to AI-outputs be difficult to get along with because they think they are wise, without being so?

## Concerns in Plato's *Phaedrus*—AI Parallels?

Socrates: "... writing shares a strange feature with painting. The offsprings of painting stand there as if they are alive, but if anyone asks them anything, they remain most solemnly silent. The same is true of written words." (275d)

- AI-5: Is it a concern w.r.t. AI that the responses one receives are static and don't (as it were) talk with us?
- This seems to be why AI applications have been developed to involve multiple steps of Q&A.



## Concerns in Plato's *Phaedrus*—AI Parallels?

Socrates: “You’d think they were speaking as if they had some *understanding*, but if you question anything that has been said because you want to learn more, it continues to signify just that very same thing forever.” (275d-e)

- AI-6: AI-output suggests that it is “spoken language,” spoken with understanding; but it isn’t.

## Concerns in Plato's *Phaedrus*—AI Parallels?

Socrates: “When it has once been written down, every discourse roams about everywhere, reaching indiscriminately those with understanding no less than those who have no business with it, and it doesn’t know to whom it should speak and to whom it should not.” (275e)

- AI-7: Is it a concern that AI-outputs don’t discriminate between users, for example, whether someone is adult or a child?

## Concerns in Plato's *Phaedrus*—AI Parallels?

Socrates: “And when it is faulted and attacked unfairly, it always needs its father’s support; alone, it can neither defend itself nor come to its own support.” (275e)

- AI-8: Is it a problem that AI-outputs cannot speak up and explain “themselves” in the face of objections?
- PS: Concerns 1-8 do not exhaust the ideas Plato discusses. We set aside a few things that require more extensive study of Plato.

## Concerns in Plato's *Phaedrus*—AI Parallels?

There is an extraordinary closeness between the lists of concerns that, in the 4th century BCE, were raised w.r.t. writing and the concerns that are raised today w.r.t. AI.

What should we make of that?

## Readings

- Chad DeChant, “Episodic memory in AI agents poses risks that should be studied and mitigated,” *Arxiv* 2025.
- Michaelian, Kourken and John Sutton, “Memory,” *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/sum2017/entries/memory/>>.
- Plato, *Phaedrus* (selection), in: ed. John Cooper, *Plato: Complete Works*, Hackett 1997.
- Connie Rosati, “The Story of a Life,” *Social Philosophy and Policy Foundation*, CUP 2013.
- Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, Armen Aghajanyan, “Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models,” *Arxiv* 2022.