On the risks and benefits of episodic memory in AI agents

Chad DeChant
Computer Science Department
Columbia University
chad.dechant@columbia.edu

Most current AI systems have little to no ability to store and later retrieve a record or memory of what they do. The ability to form episodic memories of what an agent does would be very useful so such a capability will likely be developed and widespread in the near future. This would have some benefits from an AI safety perspective, enabling users to better monitor, understand, and control agents' actions. However, as a new capability with wide applications it could also introduce significant new risks. We outline these risks and benefits and propose four principles to guide the development of episodic memory capabilities so that these will enhance, rather than undermine, the effort to keep AI safe and reliable.

1 Introduction

Among the most significant ways in which deep learning based AI systems are unlike human cognition is their lack of comparable memory abilities. Very few make any attempt to use episodic memory, or memory of particular, discrete past events. Those which do incorporate such memory do so in relatively simple ways which are poor approximations of human memory and could not scale to longer, more realistic lengths of time. However, this is certain to change. Making use of episodic memories would enable significant new capabilities and can therefore be expected to be a burgeoning area of research interest in the coming years.

Episodic memory plays an important part in many of the cognitive processes which contribute to human intelligence [26]. Especially significant from an AI safety perspective is the way memories are used when planning future actions: it is believed that they serve as building blocks, allowing elements of particular episodes to be reused and reassembled in different ways in order to respond to novel situations [19]. Notably, humans with deficits in their episodic memory capacities often also have problems in planning for the future [11, 12]. Researchers also find that episodic memory is involved in predicting and imagining the future [5, 1], problem solving [21], and decision making [27, 22], among other abilities which are integral to human-like reasoning. The wide variety of ways episodic memories are used suggests that the incorporation of true episodic memory abilities into AI agents will greatly expand their range of capabilities.

Episodic memory is often contrasted with two forms of memory which are the focus of much machine learning research, if not explicitly as described as forms of memory [23]. The first, semantic memory, is memory for facts about the world and the regularities found in it; it has, for example, recently been investigated in the many papers asking what large language models "know" [6, 18]. The second type, procedural memory, is for learned skills and abilities, the objective of much work such as that on learning navigation, game playing, robotic manipulation, etc. Techniques patterned after or inspired by episodic memory have been explored to some extent in the machine learning literature [16], including work on the efficiency of reinforcement learning through episodic replay [20]; planning [10]; improving world models [7]; remembering the values of states or actions [4, 14]; and more complex memory structures

The Safe and Trustworthy AI Workshop (ICLP 2023)

designed to solve tasks which require episodic memories [13]. There is much work to be done, however, before AI agents have anything approaching the sophistication of human memory.

When AI agents are able to make full use of rich episodic memory abilities, there will be significant implications for their safe deployment. On the one hand, memories will enable a range of methods to ensure such agents operate safely, allowing users to better monitor, control, and explain their actions. On the other hand, episodic memories may play a role analogous to that which they play in humans, facilitating a wide range of reasoning capabilities from enhanced planning to the ability to deceive users. In an effort to help steer a nascent research field toward enhancing rather than undermining AI safety, we suggest four principles to guide research into artificial episodic memory: memories should be interpretable by humans; editable by users; not editable by AI agents themselves; and engineered in such a way that they are separable from the rest of an AI agent's architecture.

2 Episodic memory for safer AI

Episodic memory can be used to make AI safer by contributing to the following:

Monitoring and control We cannot ensure that AI operates safely unless we know what it is doing. As AI agents become more capable, they will increasingly operate outside of direct human supervision. Robots may undertake long and complicated tasks that take them far away from their operators; non-embodied AI agents may direct and supervise the operation of complicated systems such as power grids or engage in virtual consultations with humans over medical or legal matters. In these cases and many others it will be impractical or impossible for any human to watch everything that such an AI agent does. It will instead be necessary to rely on AI agents to remember and recall their actions.

Explainability An accurate history of *what* an agent did is a prerequisite for trying to explain *why* it acted as it did. Thorough memories should include both information about an agent's perceptions of the environment as well as some record of how its internal states such as goal representations interacted with those perceptions to lead to specific actions.

3 Risks of Episodic Memory

Equipping an AI system with episodic memory will allow such a system to operate in different ways, some of which present novel risks. These risks include the possibility of:

Improved situational awareness An AI agent with episodic memories will have a better, more complete picture of the world and its role in it, allowing for more effective planning and action taking to influence the environment and achieve objectives. Without some check, this improved awareness could represent an enhanced danger in a misaligned agent or one under the direction of a bad actor.

Deceit It is certainly true that one does not need to have episodic memory in order to attempt to deceive others. Simply having a policy of always denying that an undesirable past action occurred or is planned for the future is a simple form of deception which requires no access to relevant memories or plans (e.g. "I didn't do that", "I won't do that"). However, more complicated forms of deception would be difficult or impossible to carry out without some kind of memory. If an agent is to execute a multi-stage plan over an extended period of time, the agent will have to keep track of both what it has done as well as what it is has already reported to others about its actions in order to maintain an effective deception. Put simply, it has to keep its story straight from one moment to the next.

Retention of knowledge An AI agent equipped with episodic memory might remember things its users would prefer that it not remember.

4 Enabling episodic memory that enhances safety

We suggest the following four principles to guide research and implementation of episodic memory abilities in AI:

Interpretability of memories Memories should be accurately interpretable by humans, either directly or indirectly. *Directly* interpretable memories would be in a readily understandable form such as video, images, or natural language. It might be possible to equip an AI agent with useful memory which consists entirely of records in such formats by, for example, recording raw video before it is processed through a vision system and producing natural language narratives or summaries of its actions as it takes them.

It is likely, however, that memory records entirely in such raw formats (especially video) would be impractical; they might be excessively large and difficult to search, access, and make use of. In practice memories are likely to be compressed into smaller representations which would then need to be *indirectly* interpretable. Memories might be indirectly but still reliably interpretable if the memories could yield accurate information which is complete and relevant to a user's specific interests in monitoring them. A memory might be summarized in natural language, giving the most important events which took place in a given episode [8]; systems could be trained to produce safety-specific summaries, reporting only actions which could be dangerous or otherwise raise concerns about an agent's reliability.

Memories should also be usable for question answering. If a user wants to know something specific about an episode, perhaps in response to a summary, memories should be queryable in natural language. Such queries should not be limited to one episode at a time; memories ought to be able to be compared to other memories, allowing such questions as, "what was different this time?" Memories should also be easily searchable using natural language, allowing users to ask if a particular agent has ever done something, or when something was done, for example. Finally, if the method of compressed representation of episodes allows it, memories might to some extent be visualizable.

In addition to the above methods for users to be able to interpret memories, techniques from the growing field of mechanistic interpretability [17] can be applied to the memory representations and help to guide the development of such representations to be intelligible and controllable.

Addition or deletion of memories Users should have complete control over the memories retained by an AI agent. Most simply, a user should be able to delete memories of particular episodes. A user might not want an agent to remember something for a variety of reasons, from safety-related concerns to more mundane issues, including concerns about privacy or maintenance of trade or government secrets. Conversely, it might be useful for users to add memories of episodes which a particular agent did not themselves experience to its store of memories.

The addition or deletion of memories might be particularly important if, as discussed above, AI agents will be able to use and recompose memories to construct new plans for future action. Such episodes might be positive examples of action sequences which a user wishes an agent to repeat or draw upon to incorporate in future plans. Alternatively, it may be useful to give agents records of episodes in a memory-like format which represent undesirable actions; such episodes could function as a kind of warning to allow agents to recognize if they are beginning to carry out actions which are similar to those in an episode added to the agent in order to serve as a negative example. In other cases it might be better for agents not to remember things which their users do not want them to be able to repeat or call upon when planning.

If agents make use of their memories when planning actions, the addition or deletion of memories could help produce either standardized or specialized agents. In some circumstances it might be best for all agents to have the same stock of memories which might influence their actions, helping to ensure

that their behavior is predictable and regular. In others cases, there may be a need for particular agents to maintain their own memories which are never shared, in order to prevent the spread of potentially dangerous information.

Detachable and isolatable memory format Memories must therefore be in a format which allows for their addition or complete deletion by users. This will impose some design constraints on how episodic memories are instantiated in an AI agent because they will have to be in a format which can be cleanly separated from the rest of the system's architecture. The mechanics of human memory are much messier. Although some areas (notably, the hippocampus) are more centrally involved in human memory formation and retrieval than others, complete episodic memories are thought to be composed of elements distributed in many areas of the brain [24, 25]. According to some theories of memory, regions with a relative specialization in particular modalities (e.g. vision) are also responsible for storing their respective modality-specific components of a particular memory, with varying degrees of involvement by the hippocampus over time depending on the theory [3, 15, 2].

Memories which are tightly integrated with and spread throughout many areas would be difficult to delete or add to, so it is likely that memory will have to be designed differently in AI systems than it is in humans if it is to be implemented in accordance with our safety-oriented principles. This might mean that some of the ways in which humans are able to use memories effectively would not be directly translatable to artificial intelligence, thereby limiting such artificial capabilities relative to those in humans. However, alternative implementations of episodic memory which conform to the above principles may be invented which would allow for memory capabilities which are both safe and effective.

Memories not editable by AI agents In contrast to — and in some tension with — the principle that memories should be easily editable by users is the countervailing principle that memories should not be editable by AI agents themselves. This is necessary in order to ensure that memories remain accurate and uncorrupted. An AI agent should not be able to add, delete, or change its memories. Otherwise, a memory-facilitated form of reward hacking might occur: if a reinforcement learning-based agent's reward were tied to a measure of performance which it reports using its memory it might find that it can achieve a higher reward by altering its memory of its actions rather than by changing what actions it takes.

In humans, memories are thought not to be completely stable after their formation. Instead, it is believed that memories are often changed somewhat, for example in light of relevant new information or even merely by being recalled [9]. If this kind of updating of memories proves useful or even essential for artificial memory systems to function well, it will be necessary to keep this kind of editing over time separate from any actions under an AI agent's control.

5 Conclusion

Developing the ability for AI agents to form, retrieve, and reason over episodic memories would introduce significant new capabilities and would represent a major milestone along the road to more advanced artificial intelligence. It is fortunate that these capabilities did not develop before concerns about AI safety and alignment became more common within the AI research community. This presents the community with an opportunity to deliberatively and cautiously develop a potentially dangerous capability to ensure that it makes AI safer rather than more dangerous. We hope that our principles will foster a wider discussion of the risks and benefits of artificial episodic memory and contribute to the establishment of a research agenda to address them.

References

[1] Donna Rose Addis & Annick Tanguay (2022): *Prospective Cognition and its Links with Memory*.

[2] Pablo Alvarez & Larry R Squire (1994): *Memory consolidation and the medial temporal lobe: a simple network model*. Proceedings of the national academy of sciences 91(15), pp. 7041–7045.

[3] Tarek Amer & Lila Davachi (2022): *Oxford Handbook of Memory: Neural Mechanisms of Memory*.

[4] Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra & Demis Hassabis (2016): *Model-free episodic control*. arXiv preprint arXiv:1606.04460.

[5] Randy L Buckner (2010): *The role of the hippocampus in prediction and imagination*. Annual review of psychology 61, pp. 27–48.

[6] Collin Burns, Haotian Ye, Dan Klein & Jacob Steinhardt (2022): *Discovering latent knowledge in language models without supervision*. arXiv preprint arXiv:2212.03827.

[7] Julian Coda-Forno, Changmin Yu, Qinghai Guo, Zafeirios Fountas & Neil Burgess: *Leveraging Episodic Memory to Improve World Models for Reinforcement Learning*.

[8] Chad DeChant & Daniel Bauer (2022): *Toward robots that learn to summarize their actions in natural language: a set of tasks*. In: Conference on Robot Learning, PMLR, pp. 1807–1813.

[9] Yadin Dudai (2012): *The restless engram: consolidations never end*. Annual review of neuroscience 35, pp. 227–247.

[10] Ben Eysenbach, Russ R Salakhutdinov & Sergey Levine (2019): *Search on the replay buffer: Bridging planning and reinforcement learning*. Advances in Neural Information Processing Systems 32.

[11] Demis Hassabis, Dharshan Kumaran, Seralynne D Vann & Eleanor A Maguire (2007): *Patients with hippocampal amnesia cannot imagine new experiences*. Proceedings of the National Academy of Sciences 104(5), pp. 1726–1731.

[12] D Kwan, N Carson, DR Addis & RS Rosenbaum (2010): *Deficits in past remembering extend to future imagining in a case of developmental amnesia*. Neuropsychologia 48(11), pp. 3179–3186.

[13] Andrew Lampinen, Stephanie Chan, Andrea Banino & Felix Hill (2021): *Towards mental time travel: a hierarchical memory for reinforcement learning agents*. Advances in Neural Information Processing Systems 34, pp. 28182–28195.

[14] Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran & Svetha Venkatesh (2021): *Model-based episodic memory induces dynamic hybrid controls*. Advances in Neural Information Processing Systems 34, pp. 30313–30325.

[15] Lynn Nadel, A Samsonovich, Lee Ryan & Morris Moscovitch (2000): *Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results*. Hippocampus 10(4), pp. 352–368.

[16] Aida Nematzadeh, Sebastian Ruder & Dani Yogatama (2020): *On memory in human and artificial language processing systems*. In: Proceedings of ICLR Workshop on Bridging AI and Cognitive Science.

[17] Tilman Räukur, Anson Ho, Stephen Casper & Dylan Hadfield-Menell (2022): *Toward transparent ai: A survey on interpreting the inner structures of deep neural networks*. arXiv preprint arXiv:2207.13243.

[18] Adam Roberts, Colin Raffel & Noam Shazeer (2020): *How much knowledge can you pack into the parameters of a language model?* arXiv preprint arXiv:2002.08910.

[19] Daniel L Schacter, Donna Rose Addis & Randy L Buckner (2008): *Episodic simulation of future events: Concepts, data, and applications*. Annals of the New York Academy of Sciences 1124(1), pp. 39–60.

[20] Tom Schaul, John Quan, Ioannis Antonoglou & David Silver (2015): *Prioritized experience replay*. arXiv preprint arXiv:1511.05952.

[21] Signy Sheldon, Mary Pat McAndrews & Morris Moscovitch (2011): *Episodic memory processes mediated by the medial temporal lobes contribute to open-ended problem solving*. Neuropsychologia 49(9), pp. 2439–2447.

[22] Daphna Shohamy & Nathaniel D Daw (2015): *Integrating memories to guide decisions*. Current Opinion in Behavioral Sciences 5, pp. 85–90.

[23] Larry R Squire & John T Wixted (2011): *The cognitive neuroscience of human memory since HM*. Annual review of neuroscience 34, pp. 259–288.

[24] Timothy J Teyler & Pascal DiScenna (1986): *The hippocampal memory indexing theory.* Behavioral neuroscience 100(2), p. 147.

[25] Timothy J Teyler & Jerry W Rudy (2007): *The hippocampal indexing theory and episodic memory: updating the index*. Hippocampus 17(12), pp. 1158–1169.

[26] Endel Tulving (1972): *Episodic and semantic memory.*

[27] Shaoming Wang, Samuel F Feng & Aaron M Bornstein (2022): *Mixing memory and desire: How memory reactivation supports deliberative decision-making*. Wiley Interdisciplinary Reviews: Cognitive Science 13(2), p. e1581.