Memory Loss in Epistemic Logic using Multi-agent Systems

Kevin Li

December 18, 2020

A common criticism in the realm of epistemic logic is it's idealization of epistemic agents. In other words, epistemic logic is only able to model the knowledge of perfect agents, which diverges in many ways from how real-life agents operate. This paper proposes a way to use the tools of distributed knowledge in multi-agent epistemic logic as a way to de-idealize epistemic logic with respect to memory. To give a clearer sense of the motivation behind de-idealizing, I begin by first discussing what exactly these idealizations are– specifically, the notion of perfect recall. With a clearer sense of what a better system should be able to express, I then lay out a way to leverage multi-agent epistemic logic to achieve this. It involves thinking about a single agent's knowledge not as just a single entity, but as a collection of instantaneous versions of the agent at different points in time. I more formally introduce a new version of the distributed knowledge operator, and a hyperparameter m, which represents a range of time that a person has memory of. I will describe several ways one might use this m parameter to represent types of epistemic agents. The second half of the paper addresses three concerns relating to intuitive aspects of memory loss that this system fails to properly capture, and proposes ways to deal with them.

1 Assumptions of perfect recall in epistemic logic

One of the ways in which epistemic logic fails to capture the knowledge of realistic agents is the assumption that epistemic agents have perfect recollection of the past, or more specifically, the knowledge that they were given in the past. This is the notion of perfect recall, which (Yap 2014) explains, ”in effect ensures that agents remember all that has previously happened to them. This means that they remember the sequence of events that has lead to the present moment, as well as their previous information states.” For an ideal epistemic agent, it's the case that when they are given knowledge, that knowledge is always available, representing information that is accessible at any point in time. For non-ideal agents, this may not necessarily be the case– they might forget information that they were given previously, or be unable to recall it at a later point in time.

For instance, consider how one might use an epistemic model to understand the information that a doorman has about the residents in a large apartment building. Suppose we want to model his knowledge of who lives in what apartment number, based on the information he gets from handing his residents packages everyday. If he has perfect memory, he is a great candidate for an agent in a typical epistemic model– every time a resident asks for a package for their apartment number, the doorman receives information that will help him make deductions in the future about who lives where. If he never forgets a face, the more packages that he hands out, the more knowledge he will have about who lives in what apartment building. However, there are some realistic circumstances that affect his knowledge– some residents take business trips, so he only sees them every few months, and some residents only pick up packages during the holidays, so he only interacts with them once a year. Given these conditions, the doorman might at one point know a particular resident's apartment number, but when asked a year later, he will probably forget. The epistemic model can no longer tell us what knowledge the doorman has, because the model takes advantage of all the information the doorman acquired in the past, but the doorman himself realistically can't recall all of it.

One way to read this concern is that there is a distinction between the kind of knowledge that the doorman has at a particular moment when he is handing over a package, versus the kind of knowledge the doorman has when considering all the package deliveries that he remembers. In other words, there is a difference between instantaneous knowledge, which is the information he has at any given moment, and recollected knowledge, which is more broadly the information he has over an interval of time. The first kind is not affected by memory loss, while the second kind is. If we can find a way to separate those two ideas in an epistemic model, it may be able to better articulate the kinds of knowledge that are relevant to non-ideal epistemic agents.

As a note here, I've made only one conceptual articulation of perfect recall here– namely, that the inability to capture memory loss is really just a lack of distinction between instantaneous knowledge and recollected knowledge. I choose to think of perfect recall in this way in order to motivate the next section, in which I propose a way to handle de-idealization with this kind of thinking. It will be evident later on, as I consider challenges to this system, that there are other ways of understanding memory loss, and therefore other (possibly conflicting) ways to explain how it complicates idealized epistemic logic.

2 Leveraging multi-agent epistemic knowledge

2.1 An agent in time

A way in which it is possible to create the distinction mentioned above is to relativize an agent's knowledge to the time in which it was introduced. In other words, for an agent who receives knowledge at multiple points in time, let's say t_1, t_2, t_3, it would be useful to separate out a notion of what knowledge was received specifically at an instantaneous moment t_i. A multi-agent epistemic system lends itself well to this job– instead of using the framework to model the beliefs of different agents, we can instead use it to model the beliefs of a single agent, but in different moments in time. Take the following grammar (Epistemic Logic slides from class) which generates the multi-agent language:

Given a set $At = \{p, q, r...\}$ of atomic sentences and a set $Agt = \{a, b, c, ...\}$ of agent symbols:

$$\varphi ::= p|\neg\varphi|(\varphi \wedge \varphi)|K_a\varphi$$

where $p \in At$ and $a \in Agt$. We read '$K_a\varphi$' as "agent a knows that φ."

If we instead think of multiple agents as one particular agent at different moments in time, we can offer a similar grammar for de-idealized epistemic logic:

Given a set $At = \{p, q, r...\}$ of atomic sentences and a set T of the integers between 1 and some arbitrary N:

$$\varphi ::= p|\neg\varphi|(\varphi \wedge \varphi)|L_t\varphi$$

where $p \in At$ and $t \in T$. We read '$L_t\varphi$' as "The agent learns φ at time t."

Here, instead of a set of agents who each have a knowledge operator, there are a set of timesteps that each have a "learn" operator, which represents the instantaneous knowledge gathered at t, independent of all other $t' \in T$ where $t' \neq t$.

One way to think of this that we've taken one agent and split them up between a bunch of time intervals. If we have an agent that exists from day 1 to N, we are really saying that they are made up of N "mini-agents", who each represent the agent on a different day: mini-agent 1 goes out and represents the agent on day 1, mini-agent 2 goes out on day 2, and so on. Each mini-agent sits out for the rest of the days, and only knows what they discovered on the day they went out.

N is an arbitrary hyperparameter– it represents the number of mini-agents, which really means either the length or granularity of time that the model represents. For example, one could think of each time step to represent an hour, in which case $N = 24$ can model an agent over the course of a day. For the doorman, who works everyday, $N = 365$ can model his knowledge over the course of a year, where each day represents an instantaneous moment of knowledge.

The semantics are the same as they are for multi-agent epistemic logic. The only thing that has changed is how we interpret and label the symbols. Since the knowledge operator is now only interpreted as instantaneous knowledge, the next section explains how collective knowledge can be reintroduced.

2.2 An agent through time

Now that an agent's knowledge has been divided up into their knowledge at particular moments in time, there needs to be a way to represent their collective knowledge, which is what they have learned across several moments in time. As explained earlier, this is the kind of knowledge which is affected by memory loss– a perfect agent is able to recall the knowledge they acquired at all points in time, but an imperfect agent, who forgets certain things they earlier acquired, is only able to recall the knowledge they acquired at some subset of time intervals. I believe this idea can be added similarly to the notion of distributed knowledge in multi-agent epistemic logic:

Given a set $At = \{p, q, r...\}$ of atomic sentences and a set T of the integers between 1 and some arbitrary N:

$$\varphi ::= p|\neg\varphi|(\varphi \wedge \varphi)|L_t\varphi|K_t\varphi$$

where $p \in At$ and $t \in T$. We read '$K_t\varphi$' as "The agent knows φ at time t." With a hyperparameter m, the truth clause for K is given by:

$$\mathcal{M}, w \models K_t\varphi \text{ iff for all } v \in W\text{: if } wR_{t'}v \text{ for all } t' \in T \text{ where } t - m < t' \leq t\text{, then } \mathcal{M}, w \models \varphi$$

This is defined in the same way distributed knowledge is, but rather than looking at a particular set of agents' distributed knowledge, we are instead concerned with a set of re-collectable moments of time, from time t. The set under consideration is defined by a parameter m, which is meant to denote the memory of an agent, or within how many time steps of time t an agent can remember. Under this representation, the set which with we ask for distributed knowledge represents some interval of time in the past of t. The intuitive picture, from the earlier idea of "mini-agents," is that we are pooling together the knowledge of each of the mini-agents that are within m days from when we ask for the agent's knowledge. If $m = 2$ and we ask about the truth of a statement $K_4\varphi$, we asking if the mini-agent who represented the agent on day 4, along with the mini-agent who represented the agent on day 3, having pooled their knowledge, can know φ. If they do, then the agent knows φ on day 4, assuming a memory capacity of 2 days.

To say a bit more on the parameter m– this can be used to model a variety of agent types. Setting $m = 1$ allows us to represent a memory-less agent, or in other words, an agent whose knowledge on a particular day is only whatever information they have learned that day. Setting $m = N$ allows us to represent a perfect recall agent, an agent whose memory extends to all moments in time in the model. The parameter can itself be a function of a point in time, $m = f(t)$, where an agent's capacity for remembering past information is dependent on the day of asking. This can be used to model an agent whose memory and therefore knowledge recollection deteriorates over time.

2.3 Example

Figure 1 is a simple example to help illustrate how this interpretation of multi-agent of epistemic logic allows us to model memory-loss, where a typical epistemic model could not. Take the doorman, who, over the course of three days, gathers information about a resident whose apartment number is some integer between 2 and 10. On Day 1, he learns that the apartment is on the East side of the building, on Day 2, he learns that the apartment is even-numbered, and on Day 3, he learns that the apartment number is either 6 or 10. One might ask if the doorman, on Day 3, knows that the apartment number is 10.

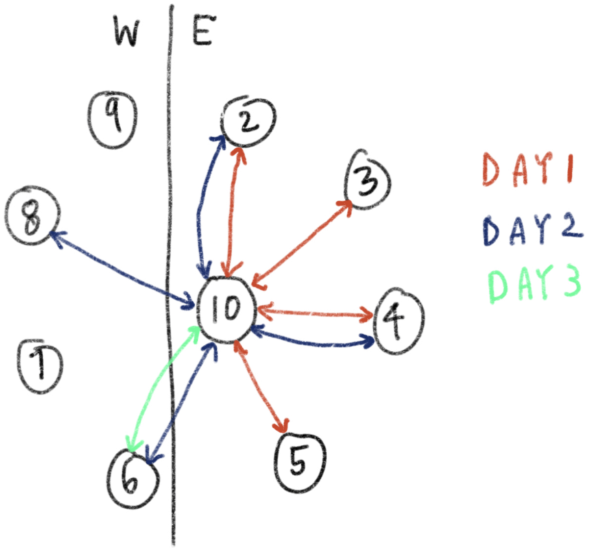

Figure 1: N = 9, m = 2. Assume reflexivity, symmetry, and transitivity.

According to the perfect recall version of the doorman, yes– over the past three days, he has been given enough information to know that the apartment number is 10. However, a more nuanced question might be: Given that the doorman is forgetful and can only remember what happened yesterday, does he know that the apartment number is 10? According to the de-idealized forgetful doorman, who has forgotten that the apartment must be on the East side of the building, the apartment can still be either 6 or 10 (the distributed knowledge between day 2 and day 3 don't give us enough conclusive information). By separating the doorman's knowledge into the days that knowledge was acquired, it's possible to ask questions that feature the idea of memory loss. The next section discusses which ideas of memory loss are still left out of the picture.

3 Other considerations for non-ideal agents

Although the system defined above is capable of representing de-idealized knowledge in certain respects that a perfect recall agent cannot, there are still other ways in which the solution is idealized.

3.1 Discontinuous memory

Suppose the doorman is great at recalling information he received a long time ago, but is forgetful in a different respect– he went out drinking at the end of day 1, and, being hungover all day on day 2, he couldn't recall what the resident told him that day. Therefore, he has knowledge of day 1 and day 3, but no knowledge of day 2. This kind of forgetfulness is different that the kind modeled above because the agent doesn't lack the ability to recall moments in time because they happened too long ago. Rather, the agent undergoes some kind of memory loss *in between* moments that they could otherwise recall. By defining the truth of the distributed knowledge operator in terms of some parameter m, we have made the idealizing assumption that agents' capacity for memory recollection falls in a certain continuous interval of time, and that it is not possible for agents to fail to remember things acquired within that interval of time.

This concern complicates the way that knowledge is defined in section 2, but not in a way that cannot be fixed. We might instead decide to define the distributed knowledge operator, L_t, not by the time of asking t and the memory parameter m, but with some set of timesteps, similar to how the original distributed knowledge operator uses a subset of agents. Rather than fixing the set to be the timesteps that fall inside a particular interval, we can define it to be *any* subset of timesteps between 1 and N. For the doorman example, using the original operator $L_3\varphi$ with $m = 2$ will take the distributive knowledge of "mini-agents" from the set $\{2, 3\}$. If we defined the operator using a set instead of t and m, we can use $L_A\varphi$ where $A = \{1, 3\}$ to denote the collective knowledge based on information gathered in the days excluding day 2. The truth condition will still behave similarly, taking the distributed knowledge over a set of timesteps.

3.2 Knowledge spillover

Another assumption that is made by this system is that agents can recollect pieces of instantaneous information, independently of the timesteps around it. However, a more intuitive perspective is that information recollected might not be what was learned on any particular day, but some inference that was made on that day with the new information learned.

For instance, suppose an agent learns φ on a given day n, but because on they had knowledge of what happened earlier that week, they infer another statement, ψ, which is a more intuitive version of φ that can be concluded based on φ and information from days $n-1$ and $n-2$. Awhile later, at a time where they can recall day n but no longer days $n-1$ and $n-2$, they accidentally recall ψ in place of φ. Another way of looking at this is that on day n, $L_n\varphi$, but because of what they remember from days $n-1$ and $n-2$, $K_n\psi$. Awhile later, when days $n-1$ and $n-2$ are no longer in memory, they recall that ψ was learned that day, even though they technically only learned φ on day n. At that later point, the information recalled on day n is not just the information on exactly day n, but to some extent influenced by the information from $n-1$ and $n-2$, even if they can't remember exactly what they learned on $n-1$ and $n-2$.

This intuitive take on memory recollection is not well-captured by a definition of knowledge based on instantaneous time steps. In other words, memory is more blurry and often imprinted based on inferential thinking with what can be recalled at a given moment, even if not all the pieces can be recovered at a later time. One way of resolving this is to argue that if what information is recalled on a day is not what was *learned* that day but what was *known* (L versus K operator,) it can only mean that they have recovered more knowledge than what they should have been able to on that day. The memory recalled can only improve if what they recalled involves information from other days that they otherwise cannot specifically recall.

3.3 Memory loss of information

As mentioned at the end of section 1, this de-idealized model of epistemic knowledge rests on the idea that the flaw of perfect recall is that it fails to distinguish between instantaneous and collective knowledge. In other words, the problem being isolated is that knowledge should be relativized to time, in order to capture the fact that memory can only be recovered from certain moments in time. A damning problem for this way of thinking is the intuitive notion that when people are unable to recall information, they are not unable to recall the time in which the information was acquired, but rather the information itself. For example, if I forget where I put my keys, I may not necessarily know at what point I had misplaced them, just that I did. The nature of my forgetfulness is not of a particular time of information acquisition, but of a particular fact, regardless of when in time I acquired it.

Of the three discussed, this concern tackles an idea of memory loss that is most divergent from the kind discussed, and requires looking at de-idealization in a different way. However, it is not necessarily incompatible with a time-based representation. Although it's unclear exactly what a fact-based memory system will look like, it may be fair to argue that such a system may still benefit from being relativized by time.

4 Concluding Thoughts

This paper offers one way to approach the shortcomings of epistemic logic as it relates to perfect recall agents. I have proposed a way of separating an agent's knowledge into their instantaneous acquisition of knowledge, and their recollection based knowledge, the former of which does not concern memory loss, the second of which does. I outline what this system may look like using the tools of multi-agent epistemic logic and the notion of distributed knowledge, and consider several assumptions made about the nature of memory loss, and how they might conflict against other intuitive approaches.

References

[1] Yap, Audrey. "Idealization, epistemic logic, and epistemology." *Synthese* 191, no. 14 (2014): 3351-3366.