Automated Commonsense Reasoning About Human Memory

Reid Swanson and Andrew S. Gordon

Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292
reid@reidswanson.com, gordon@ict.usc.edu

Abstract

Metacognitive reasoning in computational systems will be enabled by the development of formal theories that have broad coverage over mental states and processes as well as inferential competency. In this paper we evaluate the inferential competency of an existing formal theory of commonsense human memory by attempting to use it to validate the appropriateness of a commonsense memory strategy. We formulate a particular memory strategy (to create an associated obstacle) as a theorem in first-order predicate calculus. We then attempt to validate this strategy by showing that it is entailed by the axioms of the theory we evaluated. These axioms were encoded into the syntax of an automated reasoning system, which was used to automatically generate inferences and search for formal proofs.

Strategic Competency

In an effort to identify the representational requirements of human strategic planning, Gordon (2004) conducted a large-scale analysis of 372 planning strategies gathered from sources in 10 different real-world domains. This analytic approach involved authoring pre-formal representations of each planning strategy with the aim of identifying each of the concepts that would have to be formalized in order to correctly define the strategy across analogous planning cases. Of the 988 unique concepts that were identified in this work, two-thirds dealt with the mental states and processes of people. Gordon (2002) organized this subset of concepts into 30 representational areas (e.g. human memory, emotions, plan following), a set which stands today as the most comprehensive characterization of the representations involved in human metacognition that is currently available.

Gordon & Hobbs (2003) began an effort to develop formal, axiomatic theories based on the concepts in these 30 representational areas to support automated inference about the mental states and processes of people. The aim of this work is to develop formal theories that achieve a high degree of *coverage* over the concepts related to mental states and processes, but that also have the necessary inferential *competency* to support automated commonsense reasoning in this domain. These formal theories are being authored as sets of axioms in first-order predicate calculus, enabling their use in existing automated reasoning systems.

The inferential competency of formal commonsense theories could be evaluated in a number of different ways. Davis (1998) suggests the use of commonsense challenge problems (e.g. Morgenstern, 2001). However, for the formal theories developed by Gordon & Hobbs based on an analysis of strategic planning knowledge, the most appropriate evaluations of inferential competency will determine if, in fact, they achieve the requirements of human strategic planning. That is, the correctness of a particular planning strategy should follow from inferences that can be made by the underlying theories used in its formal representation.

In this paper we evaluate the inferential competency of a formal theory of mental states and processes by attempting to use it to validate the appropriateness of a commonsense planning strategy. Specifically, we attempt to use Gordon & Hobbs theory of human memory (2003) to generate a formal proof of the correctness of a human memory strategy for remembering to do things at certain times, namely "*Create an associated obstacle*". We define the correctness of this strategy as a theorem, and employ an automated theorem-proving application to evaluate the competency of the theory of human memory in generating its formal proof.

A Formal Theory of Human Memory

The representational area of human memory concerns the concept of memories in the minds of people that are operated upon by memory processes of storage, retrieval, memorization, reminding, and repression, among others. The formal theory of commonsense human memory presented by Gordon & Hobbs (2003) attempts to support inference about these processes with encodings of roughly three-dozen memory axioms in first-order predicate calculus. Key aspects of this theory can be characterized as follows:

1. *Concepts in memory*: People have minds with at least two parts, one where concepts are stored in memory and a second where concepts can be in the focus of one's attention. Storage and retrieval involve moving concepts from one part to the other.

2. *Accessibility*: Concepts that are in memory have varying degrees of accessibility, and there is some

threshold of accessibility for concepts beyond which they cannot be retrieved into the focus of attention.

3. *Associations*: Concepts that are in memory may be associated with one another, and having one concept in the focus of attention increases the accessibility of the concepts with which it is associated.

4. *Trying and succeeding*: People can attempt mental actions (e.g. retrieving), but these actions may fail or be successful.

5. *Remember and forget*: Remembering can be defined as succeeding in retrieving a concept from memory, while forgetting is when a concept becomes inaccessible.

6. *Remembering to do*: A precondition for executing actions in a plan at a particular time is that a person remembers to do it, retrieving the action from memory before its execution.

7. *Repressing*: People repress concepts that they find unpleasant, causing these concepts to become inaccessible.

Some of the axioms that are defined in this theory include predicates that are defined elsewhere, including aspects of temporal relations (Hobbs, 2002) and causality (Hobbs, 2001). Other predicates in this theory are undefined, as they do not appear in any published work.

We selected the automated reasoning engine OTTER from Argonne National Labs (Kalman, 2001) as our evaluation platform. Accordingly, each of the axioms presented by Gordon & Hobbs were translated into the first-order predicate calculus syntax that OTTER accepts, which is then automatically converted into conjunctive normal form using quantification and skolemization.

Strategy: *Create an Associated Obstacle*

To evaluate the inferential competency of this theory of human memory in its ability to prove the correctness of strategies, we selected a single strategy that dealt specifically with the way that people manage their own memory processes. The strategy, which we refer to as "*Create an associated obstacle*", is one that is easily recognized in the following illustration:

A man has a particularly noisy dishwashing machine in his home, and prefers not to run the machine when he is around. The man has a plan, which is to turn the dishwasher on in the morning as he is leaving for work so that he has clean dishes when he returns in the evening. However, this plan fails to work on most days, as the man forgets to turn on the dishwasher in the morning as he is leaving his home. To solve this problem, the man decides to place the container of dishwasher soap at the exit of his home whenever the dishwasher is full of dirty dishes. This way, whenever he is leaving the house, he will be reminded to turn on the dishwasher when he is moving the container of dishwasher soap out of his way.

This same strategy accounts for the reason that people leave notes for themselves on their car steering wheels. Likewise, people carrying bicycles on top of their cars on roof-mounted racks hide their own garage door openers inside of a cycling helmet so that they remember to bring the bicycles down before entering the garage. In each case, failing to remember an important plan step will cause the plan to fail, so an obstacle to moving forward in the plan is created that will force the person to remember the plan step associated with the obstacle.

The strategy that this person is using can be understood only with respect to a somewhat sophisticated (yet still commonsense) model of human memory. We must imagine that before employing the strategy, the plan that this man had was failing because a step in the plan was not remembered as intended. We must consider that placing the dishwasher soap in front of the exit would cause it to be in the focus of attention for some brief amount of time, during which the associated step in the plan would be more easily remembered. This would permit the man to succeed in remembering to do the plan step, leading to the successful execution of the plan.

The formal theory of human memory authored by Gordon & Hobbs includes some axioms relevant to many of these inferences, making this strategy well suited as a first test case to evaluate the competency of the theory.

Defining the Strategy Theorem

Determining whether the inferential theory is competent enough to support human strategic reasoning can be recast as a theorem-proving problem. It is a theorem that the execution of a particular strategy leads to the successful achievement of a particular goal, and the proof of this theorem will require competency in component theories. By casting a strategy as a theorem to be proved using automated theorem-proving techniques, we can quickly identify if and where parts of the theory are inadequate.

The memory strategy of creating an associated obstacle can be formulated as a theorem that involves abstract people, plans, objects, and times. One formulation, presented here in the predicate calculus syntax of OTTER, is as follows:

```
(all person time1 time2 plan step1 step2
  (intends(plan, person)
   & includes(plan, step1, time1)
   & includes(plan, step2, time2))
   & (exists object location
       (associated(object, step1, person)
        & (at(object, location, time2) ->
             prevented(step2, time2))) ->
  ((exists step0 time0
    ((do(step0, person, time0) ->
     (at(object, location, time2)) ->
   ((includes(plan, step0, time0) ->
       (remember(step1, person, time2)))).
```

In English: If a *person* has a *plan* to do something that includes two steps (e.g. turning on the dishwasher and leaving home) at two different *times*, and there exists some *object* (e.g. dishwasher soap) that is associated with the first *step1* and would prevent the person from doing the

second *step2* if it were at some *location*, then if there exists a *step0* such that doing *step0* at *time0* leads to the object being at the *location*, then the inclusion of *step0* in the plan leads to the *person* remembering *step1* at *time2*.

However, there are a number of problems with this formulation of the strategy theorem that leaves little hope of proving the strategy using only a commonsense theory of human memory. First, it requires some treatment of intentionality with regard to plans composed of steps (e.g. if you intend a plan, you intend its parts as well). Second, it requires a treatment of space and location, and the sort of steps that can lead to objects being at locations. Third, it requires a rather sophisticated treatment of the causality involved in prevention, where actions that are prevented at one moment can be enabled if some other action is taken (like moving the dishwasher soap out of the way). Fourth, it requires some understanding of the relationship between steps that involve changing the location of things and a person's focus of attention. None of these issues are dealt with by Gordon & Hobbs' theory of human memory or its supporting theories.

A more forgiving version of the strategy theorem would simply say that having an obstacle in the focus of attention leads to the retrieval of an associated plan step. This can be encoded as follows, where the "*inm*" predicate is used here to note that an obstacle is in the focus of attention of the person identified by the "*focus*" predicate:

```
(all person focus time obstacle step
  (associated(obstacle, step, person)
   & scheduled(step, time, person)
   & focus(focus, person)
   & inm(obstacle, focus, time)) ->
  remember(step, person, time)).
```

In English: If a *person* has scheduled to do some *step* at a certain *time*, and they have in their *focus* of attention some *obstacle* that is associated with that *step*, then they will remember the *step*.

This version of the theorem lacks many of the characteristics that we would like to see in a more comprehensive evaluation of the strategic competency of formal theories in general, but it is well suited for judging the inferential competency of Gordon & Hobbs theory of human memory for this one particular strategy.

Validating the Strategy Theorem

One approach to validate this strategy is to attempt a formal indirect proof of the theorem by asserting all the necessary conditions and denying the truth of the consequent. That is, a person exists that has scheduled a step and who has in their focus a concept associated with the step, but does not remember the step. However one should immediately note that a proof of this kind will be difficult and should in fact fail. Associations between concepts, even when very strong, are not enough to guarantee that a person will remember an associated concept. It could be the case that, even though there is a strong association between dishwashing soap and dishwashers, the accessibility of the step is so low that the increase caused by thinking about the associated concept is still not enough to enable its retrieval (perhaps this person dislikes doing the dishes so much that they have repressed the plan step beyond retrieval). In any formal, inferential theory that is purely qualitative with respect to the accessibility of concepts (as in the Gordon & Hobbs theory), the strongest claim that can be made is that the strategy is more likely to succeed (the concept is more likely to be remembered) than if it wasn't followed. Concluding truthfully that the step would definitely be remembered is not possible because of the indeterminate nature of the retrieval process.

We addressed this problem by considering the conclusions that can be drawn in comparative cases. That is, we challenge the strategy theorem by encoding a world that includes two different situations. In one situation, a person who has scheduled a plan step has a concept associated with a step in their focus of attention. In the second situation (represented simply as occurring at a different set of time), this same person does not have the associated concept in the focus of attention. Different worlds are then formulated for the four possibilities with regard to the retrieval of the step in each situation, as follows:

World 1: The person remembers the step both when they are focused on the associated concept and when they are not. This is consistent if the step is naturally easy to remember. This world was encoded as follows:

```
(exists m a0 a1 obst step t1 t2 f p
  (associated(obst, step, p)
   & focus(f, p)
   & memory(m, p)
   & scheduled(p, step, t1)
   & inm(p, f, t1)
   & scheduled(p, step, t2)
   & inm(p, m, t2)
   & remember(p, step, t1)
   & remember(p, step, t2)).
```

World 2: The person does not remember the step in either situation, regardless of whether they are focused on an associated concept (both *remember* clauses are negated). This is consistent if the step was hard to remember, and focusing on the associated concept was not enough to enable retrieval.

World 3: The person remembers the step only when focused on the associated concept (on the first *remember* clause is negated). This is consistent if retrieving the step required the added accessibility gained by thinking about the associated concept.

World 4: The person remembers the step only in the case when they are not focused on the *associated* concept (only the second *remember* clause is negated). This is not

consistent (and should yield a contradiction), as focusing on an associated concept should help retrieval, not hurt it.

Given these four world formulations, our challenge was to demonstrate that the necessary inferences in each of these four cases could be automatically generated by Gordon & Hobbs theory of memory. Only in one case (the fourth world) were we attempting to achieve a proof of inconsistency.

Implementing the Formal Theory

In order to evaluate the competency of Gordon & Hobbs memory theory, we encoded each of the axioms in the theory into the first-order predicate calculus syntax supported by OTTER (Kalman, 2001). The theory of memory included a number of axioms that were not related to the validation of the particular theorem in question in any way. To minimize the search space and reduce the possibility of encoding errors, we only included in our encodings those axioms of the theory that were relevant in some way (e.g. we ignored the Freudian notion of repressing memories).

Several problems arose when encoding these axioms. First, several of the predicate forms used by Gordon & Hobbs were defined in theories outside of the one formulated for human memory (e.g. the concept of *trying* depends on a definition of *goals*). This was most problematic with respect to the essential concepts concerning causality, which were largely defined by Hobbs in other work (e.g. Hobbs, 2001). We addressed this issue by authoring roughly equivalent formulations of these memory axioms that did not rely on predicates outside the domain. This typically resulted in axioms with somewhat different semantics than was originally intended, however we believe the core meanings remained relatively unchanged.

Second, the axioms of human memory were authored in a manner that assumed the existence of some elementary relational syntax that is critical to their inference competency. In some cases, this syntax was not supported in Otter. In particular, the memory theory expresses a partial ordering of the different accessibility of concepts in memory using the less-than operator (<). We addressed the lack of support for this notation in OTTER by defining an explicit predicate for less-than orderings between accessibilities (and between accessibilities and the memory threshold).

Third, function predicates were used pervasively throughout the memory axioms, which OTTER does not distinguish between relational predicates. For example, the accessibility of a concept in memory is defined as a function `a=accessibility(p, c, t)`, where p is a person, c is a concept, t is a time and a is an accessibility value. A person's memory threshold is also defined similarly. To work within OTTER's limitations, we substituted functional predicates with relational predicates where possible, e.g. accessibility was treated as a relation between a value, a person, a concept, and a time: `accessibility(a, p, c, t)`.

In the commonsense interpretation, the accessibility of a concept in memory plays a crucial role in one's ability to remember it at a particular time. In Gordon & Hobbs' theory, a concept can only be remembered if the memory threshold of the person is less than (or equal) to the accessibility of the concept. To encode this essential axiom of the theory in OTTER, we employed both the new ordering predicate as well as accessibility relation in order to reformulate Gordon & Hobbs' axiom concerning possible retrieval, as follows:

```
(all p c t (possible_retrieve(p, c, t)
  -> (exists a m
       (accessibility(a, p, c, t)
        & mthreshold(m, p)
        & less_than(m, a))))).
```

An additional axiom of the theory that is essential in generating the appropriate inferences to support the strategy theorem concerns the role that association between concepts has in aiding retrieval. As in the original theory, thinking about a concept leads to a greater accessibility of any associated concept. However, we reformulate this axiom in a manner that avoids notions of causality and change by simply ordering the accessibilities of the associated concept, where the accessibility of the concept is greater (or the same) at times when an associated concept is in the focus (not in the memory) of attention.

```
(all m c1 c2 p f t1 t2 a1 a2
   (associated(c1, c2, p)
    & focus(f, p)
    & memory(m, p)
    & inm(c1, m, t1)
    & inm(c1, f, t2)
    & accessibility(a1, p, c2, t1)
    & accessibility(a2, p, c2, t2) ->
   (less_than(a1, a2))).
```

After encoding and debugging our axiom set, we generated inferences (and searched for one proof) in conjunction with the four existentially quantified statements that defined each of the four possible worlds. In each case, we employed the *set-of-support* search strategy (as described by Kalman, 2001), with the theory given as usable in the formula list, while the worlds were each described in the set-of-support formula list. We applied a combination of both hyper-resolution and negative hyper-resolution to generate inferences, which we determined to be more interpretable than results obtained using binary resolution.

Results

Judgments of the competency of Gordon & Hobbs' formal theory of human memory were made by examining the inferences generated for each of the four world

descriptions described above.

World 1: In the case where the step is remembered regardless of whether the associated concept is in the focus of attention, OTTER successfully terminated without a contradiction. Examining the generated inferences revealed encouraging inferences. These inferences included the fact that the person tried to remember the plan step in both cases, and that it was possible to retrieve the plan step in both cases. It was also inferred that the accessibility of the plan step was lesser when the associated concept was in memory than when it was in the focus of attention. The accessibility of the retrieved step in both situations had to be less than the memory threshold for this person.

World 2: In the case where the step is not remembered in either situation, OTTER successfully terminated without a contradiction. As in world 1, it was inferred that the person tried to retrieve the step in both situations, but the only additional inferences that could be drawn were that it was not possible to retrieve the step in either case.

World 3: In the case where the step is remembered only when the associated concept is in the focus of attention, OTTER also successfully terminates without a contradiction. Among the relevant inferences, it is determined that the step is not possible to retrieve in the case where it is not remembered, and that when the associated concept is in the focus of attention, the memory threshold of the person is less than the accessibility of the step.

World 4: In the case where the step is remembered only when the associated concept is in memory (and not in the focus of attention), OTTER again terminates without a contradiction. However, we expected that OTTER should terminate with a proof by contradiction, by reasoning as follows: In the case where the step was remembered, its accessibility must have been greater than the memory threshold. Through the application of the association axiom, it should be inferred that the accessibility of the step should have been even greater when an associated concept was in the focus of attention. Therefore, this greater accessibility should also be greater than the memory threshold of the person. However, since the person tried and failed to remember the step, it follows that the accessibility was less than the memory threshold, leading to a contradiction.

It was successfully inferred that the person tried to retrieve the step in both cases. It was possible to retrieve the step when the associated concept was in memory, and the accessibility of the step in this case was greater than the memory threshold. Likewise, it was inferred that the step was not possible to retrieve when the associated concept was in the focus of attention, but fails to infer that its accessibility is less than the memory threshold.

We tried a number of different approaches to overcoming the difficulties in generating a successful proof. We found that OTTER was more successful in generating appropriate inferences if we employed binary resolution rather than hyper-resolution and negative hyper-resolution, particularly when the axiom set is tweaked. It is then possible to infer that the accessibility of the step when the associated concept is in focus is also greater than the threshold, but fails to also infer the contradiction of this. Still, many of the crucial inferences in this proof could not be generated. We believe that these failures may be due to difficulties that we had in formulating several key properties of the theory. In particular, we are concerned with problems of formulating partial orderings, uniqueness, and relational (rather than functional) definitions of accessibilities and memory thresholds.

Conclusions

The aim of this paper was to investigate the quality of current commonsense theories of human mental states and processes. Specifically, we evaluated the competency of a formal theory of commonsense human memory (Gordon & Hobbs, 2003) by attempting to automatically prove the validity of a common human memory strategy. Our results indicate that this theory does, indeed, have the breadth of axioms necessary to infer each of the inferences required to match commonsense intuitions concerning this strategy. However, direct encodings of this theory for use in a first-order predicate calculus reasoning engine (OTTER) were difficult to author, and our attempt did not yield a system that was competent enough to prove the inconsistency of situations that are impossible from a commonsense perspective. *In theory*, the formal theory may have the necessary competency. However, *in practice*, our encoding of this formal theory does not have the necessary competency.

Our work in this area has led us to rethink the assumption that the inferential competency of a formal theory on its own can be legitimately evaluated. The assessed inferential competency of a formal theory is highly dependent on the approach used to encode it for use in an automated reasoning system, the representational choices made in formulated the theorems of evaluation, and the resolution algorithm (and parameter settings) used in controlling the inference process. A challenge for future research in this area is to develop new evaluation metrics for formal theories where enough of these factors can be held constant to enable legitimate assessments that are comparable across different reported results.

Acknowledgments

The project or effort depicted was or is sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and that the content or information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Davis, E. 1998. The Naive Physics Perplex. AI Magazine, Winter 1998.

Gordon, A. 2002. The Theory of Mind in Strategy Representations. Proceedings, Twenty-fourth Annual Meeting of the Cognitive Science Society (CogSci-2002), George Mason University, Aug 7-10. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, A. 2004. Strategy Representation: An Analysis of Planning Knowledge. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, Andrew S. & Hobbs, Jerry R. (2003) Coverage and Competency in Formal Theories: A Commonsense Theory of Memory. *Proceedings of the 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning,* Stanford University, March 24-26, 2003.

Hobbs, Jerry R. (2001). Causality. *Proceedings of the Fifth Symposium on Logical Formalizations of Commonsense Reasoning*, pp. 145-155). NYU, New York.

Hobbs, Jerry R. (2002) Towards an Ontology for Time for the Semantic Web. In the proceedings of LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language, pp 28-35. Las Palmas, Spain.

Kalman, John. A. (2001) *Automated Reasoning with Otter*. Paramus, NJ: Rinton Press.

Morgenstern, L. (2001) Mid-Sized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking, Studia Logica, vol. 67, 2001.