Formalizations of Commonsense Psychology

Andrew S. Gordon and Jerry R. Hobbs

■ The central challenge in commonsense knowledge representation research is to develop content theories that achieve a high degree of both competency and coverage. We describe a new methodology for constructing formal theories in commonsense knowledge domains that complements traditional knowledge representation approaches by first addressing issues of coverage. We show how a close examination of a very general task (strategic planning) leads to a catalog of the concepts and facts that must be encoded for general commonsense reasoning. These concepts are sorted into a manageable number of coherent domains, one of which is the representational area of commonsense human memory. We then elaborate on these concepts using textual corpus-analysis techniques, where the conceptual distinctions made in natural language are used to improve the definitions of the concepts that should be expressible in our formal theories. These representational areas are then analyzed using more traditional knowledge representation techniques, as demonstrated in this article by our treatment of commonsense human memory.

Among the more challenging problems in the field of artificial intelligence are those that require computers to engage in commonsense reasoning. In representational areas where robust content theories exist, a whole suite of applications becomes possible. For example, given a commonsense ontology of time (as in Hobbs 2002), we can construct automated reasoning systems to tackle real-world problems associated with transportation logistics, event planning, and factory process scheduling that are robust in the face of real-world concerns like time zones, daylight savings time, and international calendar variations.

Given the importance of an ontology of time across so many different commonsense reasoning tasks, it is appropriate to devote effort and special attention to this representational area so as to develop an inferential basis that is logically sound. The same is true of many of the other content theories that have traditionally defined the scope of knowledge representation research, especially ontologies of events (Shanahan 1995), space (Cohn and Hazarika 2001), and physical entities (Davis 1993).

Despite the progress that has been made in engineering automated reasoning systems and expressive logical languages, the bottleneck continues to be the lack of large-scale content theories across the full breadth of commonsense representational areas. There have been significant efforts in the last few years in developing large-scale commonsense resources. Two such efforts are OpenCyc (www.opencyc.org) and the Suggested Upper Merged Ontology (Niles and Pease 2001). For the most part, however, these efforts have lacked a coherent empirical methodology for determining what content they should cover, and in part as a result of this, they are weak in areas of commonsense psychology, for example, an area that is critical for many aspects of strategic planning.

Indeed, when surveying the field of knowledge representation as a whole, one gets the sense that most knowledge representation researchers are more comfortable with concepts related to the natural sciences (for example, physics) than the social sciences (for example, psychology). Considering that tremendous progress has been made in commonsense reasoning in specialized topics such as thermodynamics in physical systems (Collins and Forbus 1989), it is surprising that our best content theories of people are still struggling to get past simple notions of belief and intentionality (van der Hoek and Wooldridge 2003). However, systems that can successfully reason about people are likely to be substantially more valuable than those that reason about thermodynamics in most future applications.

Content theories for reasoning about people are best characterized collectively as a theory of commonsense psychology, in contrast to those that are associated with commonsense (naïve) physics. The scope of commonsense physics, best outlined in Patrick Hayes's first and second "Naïve Physics Manifestos" (Hayes 1979, 1984), includes content theories of time, space, physical entities, and their dynamics. Commonsense psychology, in contrast, concerns all of the aspects of *the way that people think they think*. It should include notions of plans and goals, opportunities and threats, decisions and preferences, emotions and memories, along with all of the other mental states and processes that people attribute to themselves and others (Clark 1987).

Our contemporary understanding of commonsense psychology has been less informed by AI than by cognitive psychology, where reasoning about the mental states of other people has been studied as *theory of mind* abilities, than by AI. Developmental psychologists have noted that these abilities are strongly age-dependent (Wellman and Lagattuta 2000; Happe, Brownell, and Winner 1998) and have argued that they are central in explaining cognitive deficiencies associated with autism (Baron-Cohen 2000) and schizophrenia (Corcoran 2001). Although alternative hypotheses have been proposed (Goldman 2000), researchers have asserted that our commonsense psychological abilities are facilitated by a tacit representation-level theory of mental attitudes and behavior (Gopnik and Meltzoff 1997; Nichols and Stich 2002). This hypothesis, referred to as the *theory theory*, is very much in line with the perspective of the average knowledge representation researcher in AI, whose aim is to describe tacit representation-level theories as explicit, axiomatic theories.

With the authoring of content theories of commonsense psychology as our goal, we can begin to confront the formidable challenge of orchestrating a research effort of this scale. In drawing a parallel between commonsense psychology and commonsense physics, we run the risk of falling into the same methodological quagmire that has plagued research in commonsense physics since its inception. In his article titled the "Naïve Physics Perplex," Davis (1998) reflects on the methodological problems that have hindered progress. He argues that the goal of commonsense reasoning research is the generation of *competency theories* that have a degree of depth necessary to solve inferential problems that people are easily able to handle.

Yet competency in content theories is only half of the challenge. Commonsense reasoning in AI theories will require that computers not only make deep humanlike inferences but also ensure that the scope of these inferences is as broad as humans can handle, as well. That is, in addition to *competency*, content theories will need adequate *coverage* over the full breadth of concepts that are manipulated in human-level commonsense reasoning. It is only by achieving some adequate level of coverage that we can begin to construct reasoning systems that integrate fully into real-world AI applications, where pragmatic considerations and expressive user interfaces raise the bar significantly. A conservative commonsense reasoning researcher might argue that coverage is an additional constraint on an already difficult task and is best addressed after suitable competency theories have been put forth. We argue that unless the issue of coverage is addressed first competency theories will be intolerant of elaboration and difficult to integrate with each other or within larger AI applications.

This article presents a new methodology for authoring formal commonsense theories, based on the belief that the problems of coverage and competency should be decoupled and addressed by entirely different methods. Our approach begins by outlining the coverage requirements of commonsense theories through the analysis of a corpus of strategies. These requirements are elaborated to handle distinctions made in natural language, as evidenced through analyses of large English text corpora. We then address the specification of a formal notation (in first-order predicate calculus) and of a full axiomatic theory. The section on representational requirements of strategic planning describes the methods used to solve the coverage problem in the domains of commonsense psychology. The section on commonsense psychology in natural language elaborates on the role of natural language in refining these representations, with an example domain of the commonsense psychology of memory. The section on a commonsense theory of human memory presents a formal theory of the commonsense psychology of memory aimed at achieving a high degree of both coverage and inferential competency. The final section offers our conclusions and considers the challenges of future work in formalizing other domains of commonsense psychology.

The Representational Requirements of Strategic Planning

Gordon (2001a) noted that there is an interesting relationship between concepts that participate in commonsense psychology theories and planning strategies, the abstract patterns of goal-directed behavior that people recognize across analogous planning cases. For example, consider a strategy that a concert pianist used as an aid in memorizing complicated compositions such that they could be executed without referring to sheet music. For particularly challenging passages, the pianist explained that he would focus not on the sensations of his hands hitting the keys during practice, but rather on the visual motions he experienced by watching his hands. He reasoned that his ability to remember complex visual patterns was sometimes more effective than his motor memory, and found that if he again focused his eyes on his hands during a performance, his expectations would guide them to do the right thing. This same strategy may be applicable to workers who operate complex machinery and is even more generally applicable to any performance-based memory task that is directly perceived by the performer. Domain-specific details of any application can be abstracted away so that the description of the strategy does not refer to musical pieces, piano keyboards, or human hands. There are some concepts in this strategy that will remain essential to every instantiation of it in any planning situation. These concepts include the commonsense psychology notions of the attending to a perception, the observation of a performance, the expected pattern of perception, and the intention of memorization, among others.

Noting the remarkable conceptual breadth of planning strategies like this one, Gordon (2001b) devised a methodology for outlining the full scope of representational requirements for planning strategies. This methodology involved the collection and analysis of a large corpus of planning strategies across many different planning domains. Three hundred seventy-two strategies were collected in ten different planning domains (business, counting, education, Machiavellian politics, performance, relationships, scientific practice, warfare, and the anthropomorphic domains of animal behavior and cellular immunology). These 372 strategies were collected by analyzing texts that were encyclopedic of strategies within the domains, by interviewing domain practitioners, and through the interpretive observation of activities within domains. Preformal representations of the strategies were authored to identify the concepts that participated in all instances of each of the strategies, regardless of the specifics of the planning situation. The full set of 372 preformal strategy representations (available in Gordon 2004) required 8,844 concepts to be expressed. Removing duplicate concepts and combining synonymous terminology reduced this set to a controlled vocabulary of 988 unique concepts.

To illustrate this approach, consider the strategy of the concert pianist that was mentioned earlier, which is one of the 39 strategies analyzed from the domain of performance. The preformal representation of this particular strategy included the following clause, where capitalization and italicization demark significant conceptual terms: “The planner *Monitors* for *Perceived actions* that are *Attended* to with the goal of *Remembering*.” The 4 italicized conceptual terms here were among the 43 that were used in the preformal representation of this strategy. After combining synonymous terms used in other strategy representations, the controlled versions of these 4 terms were identified as *Monitor*, *Observed execution*, *Attend*, and *Memory retrieval*.

The ontological scope of the full set of controlled terms was very broad. To better understand this scope, we clustered the 988 terms and grouped them into 48 representational areas that corresponded to traditional areas of research in knowledge representation or cognitive science. Of the 48 representational areas, 8 of them (189 unique concepts) were closely related to fundamental topics that have been well addressed in knowledge representation research. These were the areas of Time, World states, Events, Space, Physical entities, Values and quantities, Classes and instances, and Sets. An additional 10 of the 48 representational areas (164 unique concepts) were related to the commonsense notion of *agency*, namely Agents, Agent Relationships, Communities and organizations, Resources, Abilities, Activities, Communication acts, Information acts, Agent interaction, and Physical interaction.

Table 1 outlines the remaining 30 of these 48 representational areas (from Gordon 2002). The 635 concepts in these 30 representational areas were related to the mental states and processes of people, broadly speaking. For example, the four controlled terms mentioned earlier from the piano performance strategy were clustered into the representational areas of Monitoring (*Monitor*), Observation of execution (*Observed execution*), Body interaction (*Attend*), and Memory retrieval (*Memory retrieval*). Collectively, these 30 representational areas constitute the set of concepts related to commonsense psychology that are necessary to represent adequately those that participate in strategic planning knowledge. More broadly speaking, they constitute the most descriptive formulation of the breadth of human commonsense psychological reasoning to date and identify the conceptual scope of a representational theory of mind.

The conceptual breadth of these 30 areas is significantly greater than previous work in knowledge representation concerning commonsense psychology. However, the methodology used to identify these concepts was based on the analysis of planning knowledge, not on the inferences that people draw concerning these concepts. As a result, no inferential theories to drive automated reasoning about mental states and process are produced by this approach—only an indication of the sorts of concepts that would participate in these inferential theories. While it is tempting simply to treat each of these concepts as a formal concept (for example, as a predicate in first-order logic), the nature of these terms poses a few significant problems. The conceptual specificity of the terms in an area is not uniform. An area such as goal management (referring to people's ability to select and prioritize the goals that they will attempt to pursue) includes some very general concepts among the 34 that were identified, such as the mental action of suspending the pursuit of a goal or the mental entity of the currently pursued goal. However, it also calls for more specific terms, such as the mental action of removing an auxiliary goal and the mental action of removing a knowledge goal after it has been achieved.

An even more significant problem exists when the evidence offered by strategy representation provides only a handful of terms to indicate the conceptual breadth of the representational area. This problem is best exemplified by the smallest representational area identified, *Memory retrieval,* concerning people's ability to store and retrieve information between the focus of their attention and their memory. Only three memory-related terms occurred in the strategy representations: the mental action of attempting to memorize something so that it could be retrieved from memory at a later time, the mental action of retrieving something from memory into the focus of one's attention, and the mental construct of a memory cue that is the trigger for a memory retrieval event. While it is conceivable that a formal inferential theory could be constructed from predicates based on these three concepts alone, our commonsense models of the human memory process are richer than this. In order to solve both the problems of conceptual specificity and sparse concepts, a second phase of conceptual elaboration is necessary.

Commonsense Psychology in Natural Language

The relation between the way people use language in communication and the sorts of formal representations of meaning that are employed in commonsense reasoning theories is complex. Knowledge representation researchers have generally avoided elaborating this relationship wherever possible. However, an opportunity exists for capitalizing on natural language as a resource to guide knowledge representation work. Natural language is still the most expressive means of making conceptual distinctions, and the analysis of written or transcribed natural language can greatly influence the conceptual distinctions to be made in formal commonsense theories whenever coverage is a concern. In the research described in this article, the expressive breadth of natural language was used to moderate the problems of strategy representations as the only indicator of the scope of concepts to be formalized into inferential theories.

We developed a methodology for elaborating the concepts in different representational areas and applied this methodology to the 30 commonsense psychology areas presented in the previous section. The methodology, described below, begins with the concepts grouped into a single representational area from the list of 30 above, and yields an elaborated set of concepts ready to be formalized into inferential theories. This methodology is language based, as it involves the large-scale analysis of natural language text data using tools and techniques borrowed from the field of computational linguistics. As in most large-scale knowledge representation efforts, the methodology is labor intensive and requires expertise outside of the traditional scope of knowledge representation research. Executing this methodology required the efforts of many graduate students who were studying linguistics, computational linguistics, or computer science. Typically, applying this methodology to an individual representational area required two weeks of full-time effort by a team of these graduate students. We describe this methodology below, using examples from the *Memory retrieval* representational area of commonsense psychology, one of the 30 commonsense psychology representational areas listed above.

Representational Area	Summary
1. Managing knowledge	Concepts of knowledge, belief, assumptions, justifications and the mental processes that manipulate these concepts in reasoning
2. Similarity Comparison	The mental processes of making comparisons and drawing analogies in order to find similarities and differences
3. Memory retrieval	The processes of storing and retrieving concepts from memory, and the effects of memorization and memory repression
4. Emotions	Emotional states and the processes of appraisal and coping with negative emotions
5. Explanations	The processes of generating satisfying explanations for effects that have unknown causes
6. World envisionment	Thinking about states in the world and the causal connections between them
7. Execution envisionment	Thinking about the effects that the actions would have if they were performed in the world
8. Causes of failure	The patterns of explanations that people use to explain why plans were unsuccessful
9. Managing expectations	Thinking about things that haven't yet happened, and the processes of being surprised or unsurprised if they occur
10. Other agent reasoning	The process of taking the perspective of another person in order to imagine what is going on in their mind
11. Threat detection	The intersection between expectations about what is going to happen and the goals that one wants to achieve
12. Goals	Concepts that describe desirable world states that are to be pursued or maintained
13. Goal themes	The justifications for goals that are based on the various roles that people hold in their lives
14. Goal management	The processes of prioritizing, pursuing, and abandoning the set of goals that one holds
15. Plans	Concepts that describe sets of behaviors that one imagines will lead to the achievement of a goal
16. Plan elements	Concepts that compose the behaviors in plans, including the notions of preconditions, conditionals, and iterations
17. Planning modalities	Differentiations in the way that one engages in the processing of planning for the purpose of achieving a goal
18. Planning goals	Preferences and constraints that guide the planning processes toward the creation of plans that are viewed as optimal
19. Plan construction	The process of planning by creating a new plan from scratch based on the behaviors that one knows how to perform
20. Plan adaptation	The process of planning by adapting an existing plan so that it achieve the goal in the current situation
21. Design	The process of planning where the goal is to create or configure something in the external world
22. Decisions	The concepts of a decision, choices, consequences, and the reasons for taking one course of action over another
23. Scheduling	The process of committing to execute plans in the future by placing them on an imaginary timeline
24. Monitoring	The processes of continuously or periodically focusing attention on a particular world state and waiting until something occurs
25. Execution modalities	Differentiations in the ways that people execute the plans that they have scheduled
26. Execution control	The process of turning a plan into a reality by engaging in the behaviors that were imagined would achieve some goal
27. Repetitive execution	The processes associated with executing plans that have iterative or repetitive components
28. Plan following	The processes of evaluating the progress of a plan that one is currently executing
29. Observation of execution	The processes of evaluating the progress of a plan where the plan is being executed by someone else
30. Body interaction	The mental states, processes, and phenomenon related to sensation and control information that passes between the mind and the body.

Table 1. The 30 Representational Areas of Commonsense Psychology.

The first step, *expression elicitation*, is to identify an initial set of natural language words, expressions, and whole sentences that native speakers judge to be expressive of concepts related to the given representational area (such as memory retrieval). The existing concepts in the representational area are used to help people think about the area, but the aim is to develop an initial set of expressions that is more indicative of the true breadth of expressible concepts. For example, the representational area of *Memory retrieval* initially included the concept of retrieving something from memory into the focus of one's attention, and it can be used to elicit a wide range of related English expressions:

> He *was reminded of* the time he crashed his car. The broken headlight *made him think of* when he crashed his car. He *remembered* the exact location of the car crash. He *recalled* the name of the street where it took place. Every car horn *evoked memories* of that fateful day.

Originally, this step in the methodology was completed largely by collaborative brainstorming among a handful of members of our research group and typically generated a dozen or so nouns, verbs, adjectives, and idiomatic expressions for each of the concepts in the representational area. Later, we found it more effective to hold large-group brainstorming sessions, where a dozen or more graduate students and staff members (all native English speakers) would each try to quickly come up with related words and phrases that were more linguistically creative and insightful than those that their peers produced. In this manner, a hundred or more expressions could be elicited over the course of a single large-group meeting.

The second step, *lexical expansion*, is to use the initial expressions to seed a more thorough search for related words and expressions in a range of linguistic reference resources. The labor of this step, conducted individually by the graduate students in our research group, involved looking up each of the elicited expressions in traditional dictionaries and thesauri, production dictionaries, and phrase dictionaries and recording other associated expressions to generate a large-coverage list. Particularly valuable resources included the *Longman Language Activator* production dictionary, the *Collins Cobuild Dictionary for Advanced Learners,* and Levin's description of English verb classes and alternations (1993). As an example, the initial set of expressions concerning memory was expanded to include verb phrases such as *to know by heart* and *to suppress the memory of* and related nouns such as *a hint* and *a memento*. Team members with more linguistics experience were typically more efficient at accomplishing this step, which generally yielded hundreds of expressions and language fragments per representational area.

The third step, *corpus analysis*, is to collect a large database of real examples of the use of language related to the representational area by encoding the relevant vocabulary into finite-state automata that can be applied to large text corpora. Large-scale corpus analysis has become commonplace in modern computational linguistics research, and many research groups have authored software designed to make it easy for researchers to collect instances of particular linguistic patterns by extracting them directly from textual data. Our group utilized the Intex Corpus Processor software (Silberztein 1999a, 1999b), which allowed us to author linguistic patterns as finite-state automata using a graphical user interface. To simplify the specification of patterns, we relied heavily on a large-coverage English dictionary compiled by Blandine Courtois, allowing us to specify components of our finite-state automata at a level that generalized over noun cardinality and verb inflections. For example, a single pattern for a memory retrieval expression can be described with finite-state automata of four successive transitions that handle both *made him think of* and *makes her think of* by generalizing over the verb and the pronoun. Members of our research group authored hundreds of generalized linguistic patterns during this step, one for every expression that is identified in the previous step. These were then combined into a single finite-state automaton that could be applied to any English text corpus to collect real examples of the use of these patterns. We applied each of these composite automata to 20th-century fiction and nonfiction works that we downloaded from the Project Gutenberg website (www.gutenberg.net), typically yielding hundreds to thousands of indexes per representational area per averaged-size book. Sentences containing these indexes were then compiled into a list (concordance) for review. Gordon et al. (2003) evaluated the quality of the finite-state automata resulting from the work involved in this step according to traditional information retrieval standards. Results indicated that this approach was effective at identifying 81.51 percent of the expressions associated with these concepts in English written text (recall score) and that 95.15 percent of the identified expressions would be judged as appropriate by a human rater (precision score).

The fourth step, *model building,* is to review the results of the corpus analysis step to identify the conceptual distinctions made in real language use. The aim of this step is to identify a set of conceptual primitives to be used in an axiomatic theory that is of broad enough coverage to capture the distinctions that are evident in the concordance. This set will serve as a replacement for the initial list identified through strategy representation. The task in this step is to cluster the sentences in the concordance by hand into sets of synonymous uses of the expressions. In our working group, this step was the only step that was not conducted by graduate students, as it relied more heavily on familiarity with the practices of formalizing knowledge domains. While it is often argued that there are no true sets of synonymous expressions, an effort was made to identify distinctions that will play functional roles in formal inferential theories of reasonable complexity. For example, there are shades of semantic difference between uses of the phrases *repression of memory* and *suppression of memory*, but we felt that it was unlikely that formal inferential theories of commonsense memory retrieval would be able to define or capitalize on these differences in the near future, so instances of the two uses were judged synonymous to the mental event of causing a concept in memory to become inaccessible.

The model-building step for the area of memory retrieval resulted in 12 clusters of synonymous linguistic uses of the expressions, which can be described as follows. People have a *memory ability* (1) that allows them to move *memory items* (2) in and out of the focus of their attention, unless they are *repressed memory items* (3). People have some intentional control over their memory, including the operators of *memory storage* (4) for memorizing things and *memory retrieval* (5) for recalling things into focus. The second operator can fail, however, resulting in a *memory retrieval failure* (6). There are unintentional actions of memory as well, such as making a memory inaccessible through *memory repression* (7). The everyday unintentional function of memory is simply to cause a *reminding* (8), particularly when some other mem*ory cue* (9) is the focus of attention. This plays a special role in the processes that surround plan execution, where you may *schedule a plan* (10) with the intention of remembering to do it at a certain time in the future, specifically during the event of a *scheduled plan retrieval* (11). But sometimes this can fail, yielding a *scheduled plan retrieval failure* (12).

A Commonsense Theory of Human Memory

Having identified a set of representational constructs that will participate in any commonsense theory of memory of broad coverage, we can now employ more traditional knowledge representation methods of formalization and axiomization. This section presents the results of applying these methods to author a theory that achieves the identified broad coverage requirements along with the necessary competency to support human-level inferences about memory.

Concepts in Memory

To remember something is to go through a change of state in which you come to think consciously of something that you knew before but were not consciously thinking of. If we are going to build an underlying theory of mind that will support and give coherence to the notion of "remembering," we must provide the language in terms of which the "before" and "after" of remembering can be described. Memory is commonly understood in terms of a spatial metaphor. There is a region in the brain that is the focus of attention, and there is a region that is the memory.

But first we need to specify what kind of entity is *in* each of these regions. We will call these entities *concepts* and say very little else about them. This is intended to be a very general term, covering several more specific terms in other subtheories. For example, a theory of belief may deal with propositions and predicates; we can stipulate each of these to be examples of concepts. A theory of perception may deal with images, and we can call these "concepts" as well. Then any statements we make here about the behavior of concepts in the mind will be true of propositions, predicates, and images.

In our theory of memory, we will assert that a person, or perhaps another agent, has a *mind.* The mind has at least two parts, or regions: a *focus* of attention (we will hereafter call this the "focus") and a *memory*. Concepts can be in the mind, and if they are, they are either in focus or in the memory. (See figure 1.) We can call the relation of a concept being in one region or another the *inm* relation, for "in-mental" or "in mind." A concept can be "inm" focus or memory for an interval of time.

The notion that a concept is in the focus of attention is intended to correspond to when one is consciously thinking of the concept. Here we will not attempt to explicate the notion of consciousness further, except insofar as

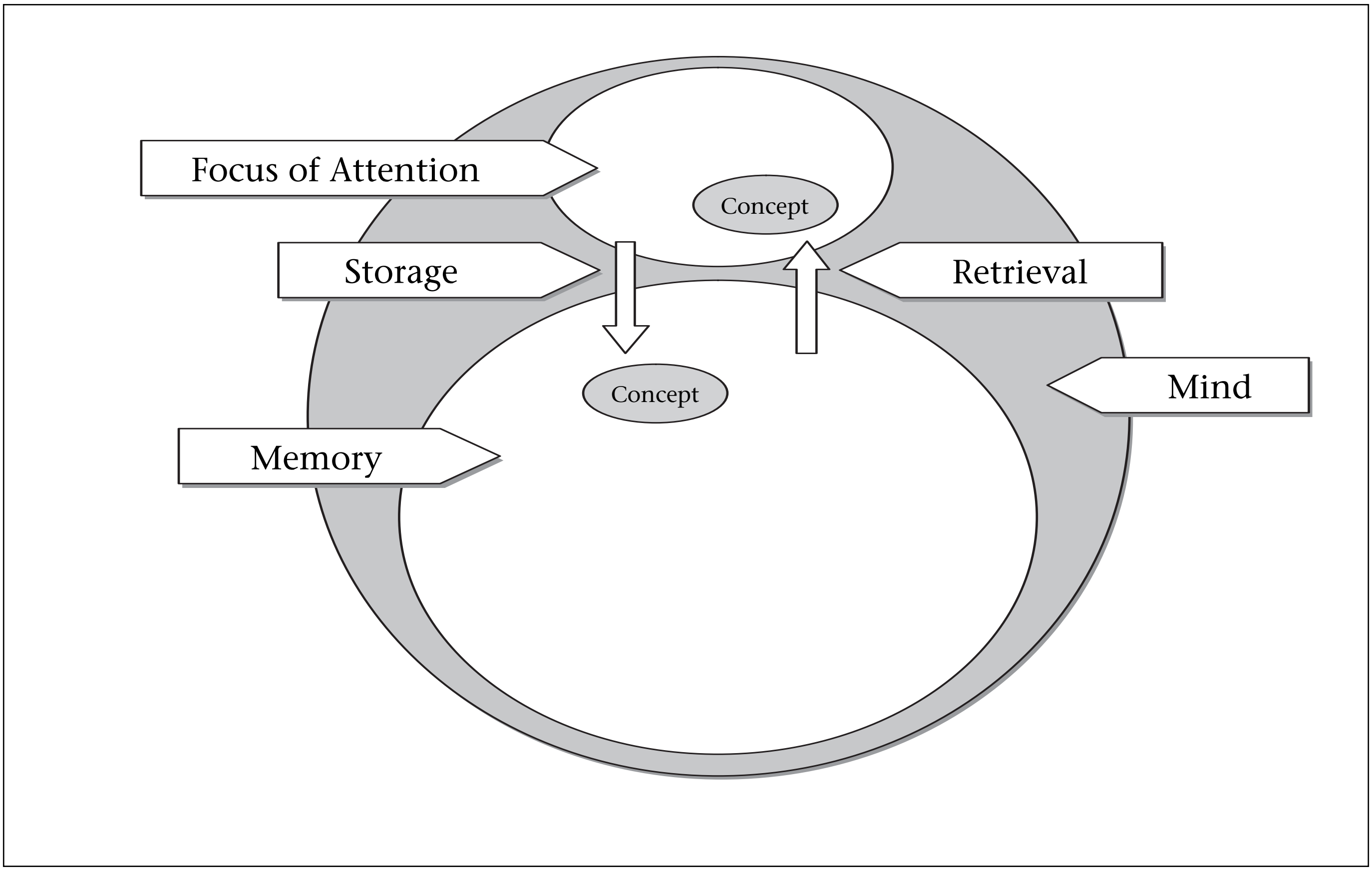

Figure 1. Memory and the Focus of Attention.

certain properties of consciousness are relevant to memory. But one could develop a much richer theory of conscious thought that captures more of its commonsense properties. In fact, in other work, we have done just that for one aspect of conscious thought—envisioning the causal consequences of actions and events in the world.

An agent *stores* a concept in memory when there is a change from a state in which the concept is in the agent's focus but not in the memory to one in which it is in the memory. The concept may or may not still be in focus.

Similarly, to *retrieve* a concept from memory is to change from a state in which the concept is in memory and not in focus to one in which it is still in memory but also in focus.

The actions *store* and *retrieve* are relations between agents and concepts at particular times.

The only way for a concept to get into an agent's memory is for it to be stored. (This rules out preexisting Platonic ideals in the memory, as in Plato's *Meno*.) Moreover, thinking about something—having the concept in focus—is a prerequisite for having it in memory.

Accessibility

We also want to capture notions like "trying to remember," "having difficulty remembering," and "failing to remember." To support ideas like these, we need to introduce further structure in our theory of concepts in memory. We need the idea that some concepts in memory are more accessible than others.

Thus, concepts in memory at a particular time have an *accessibility*, which is an element in a partial ordering. *Accessibility* is a function mapping a concept, an agent's memory, and a time into an element of the partial ordering.

Accessibility may or may not be comparable across agents. You may have an easier time remembering, say, the laws of elementary mechanics, than others do, but normally we could talk of that in behavioral terms.

Once we have accessibility, we can talk of inaccessibility as well. We often have the experience of not being able to remember something at one particular time and being able to remember the same thing a little later. Sometimes the accessibility of a concept is so low, the concept is inaccessible.

Thus, for any given agent, there is an accessibility value (in the partial ordering) below

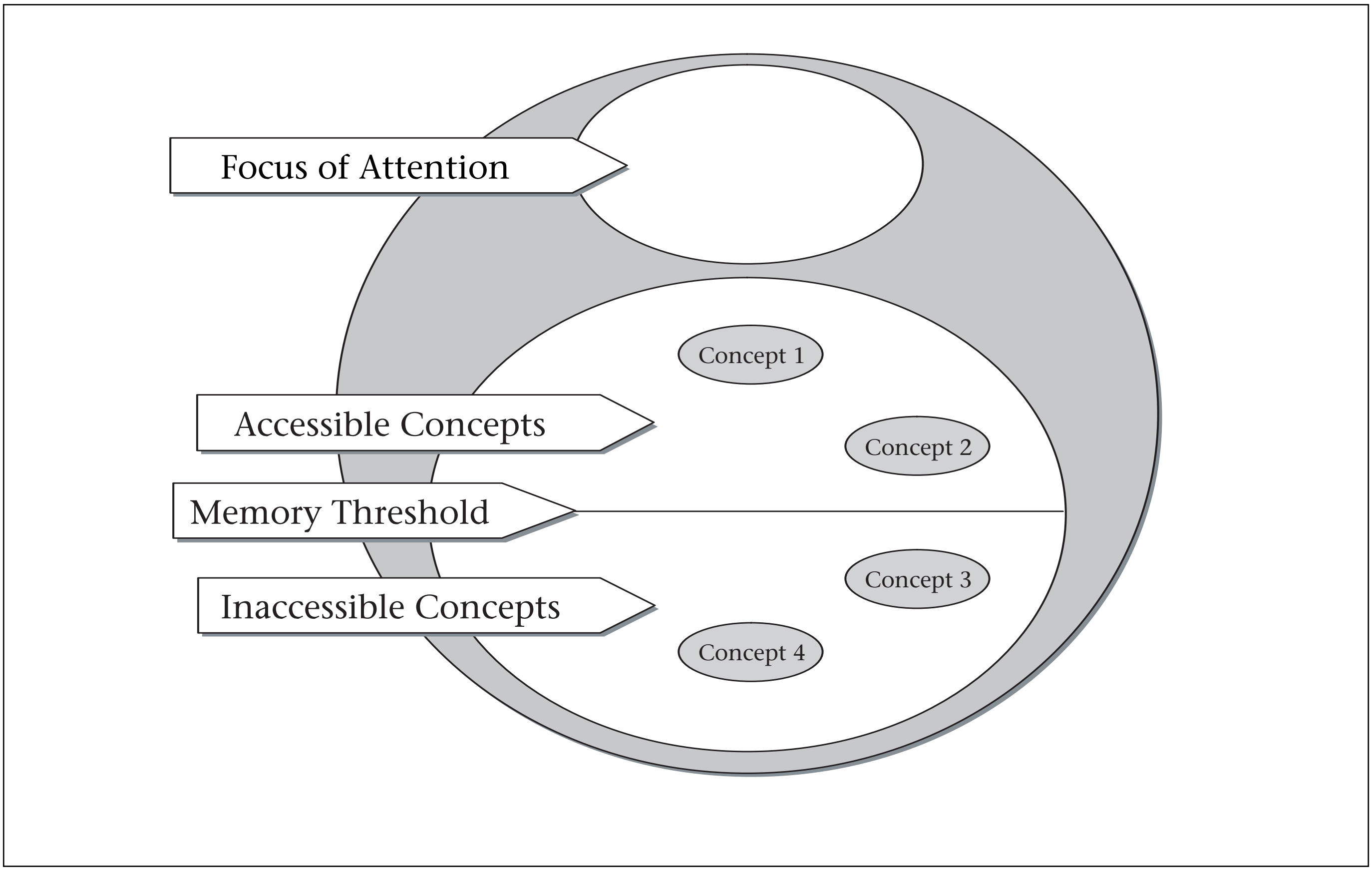

Figure 2. The Memory Threshold.

which concepts cannot be retrieved from memory. We can call this the "memory threshold," or *mthreshold*. (See figure 2.)

Our theory is silent on how long a particular concept retains a particular accessibility value. After all, that is one of the mysteries of human existence and would not be a part of any commonsense theory. However, we can explicate some features of the causal structure underlying changes in accessibility, and we do that in the next section.

Associations and Causing to Remember

One concept can *remind* an agent of another concept. This occurs when the first concept being in focus causes the second to be in focus as well. If our theory of memory is going to support a notion of "reminding," we will need an account of how this could happen.

One concept can be *associated* with another for a given agent. The specific kinds of association could be partially explicated in a theory of the structure of information or other commonsense theories. For example, inferentially related concepts are associated for agents that know the inferential relations. Parts are associated with wholes, and instruments are associated with the actions they play a role in. Two concepts might be associated for an agent for a completely random reason; someone might associate the color red with his or her first-grade teacher. Associations between concepts might be set up on a temporary and ad hoc basis; for example, someone might associate a string around his or her finger with the action of buying toothpaste at the grocery on the way home from work on a particular day.

In our theory of memory we do not attempt a deeper analysis of the idea of association; rather we concern ourselves with the causal consequences of concepts being associated. In the current theory we take associations to be dependent on agents but not on times, although times could easily be incorporated.

The basic fact about association in the theory of memory is that if you remember something, that will raise the accessibility of the concepts that are associated with it. (See figure 3.)

All of the predicates and functions we have introduced so far—*focus*, *memory*, *inm*, *store*, *retrieve*, *accessibility*, *mthreshold*, and *associated*—have been explicated in our theory of memory in axioms in first-order predicate cal-

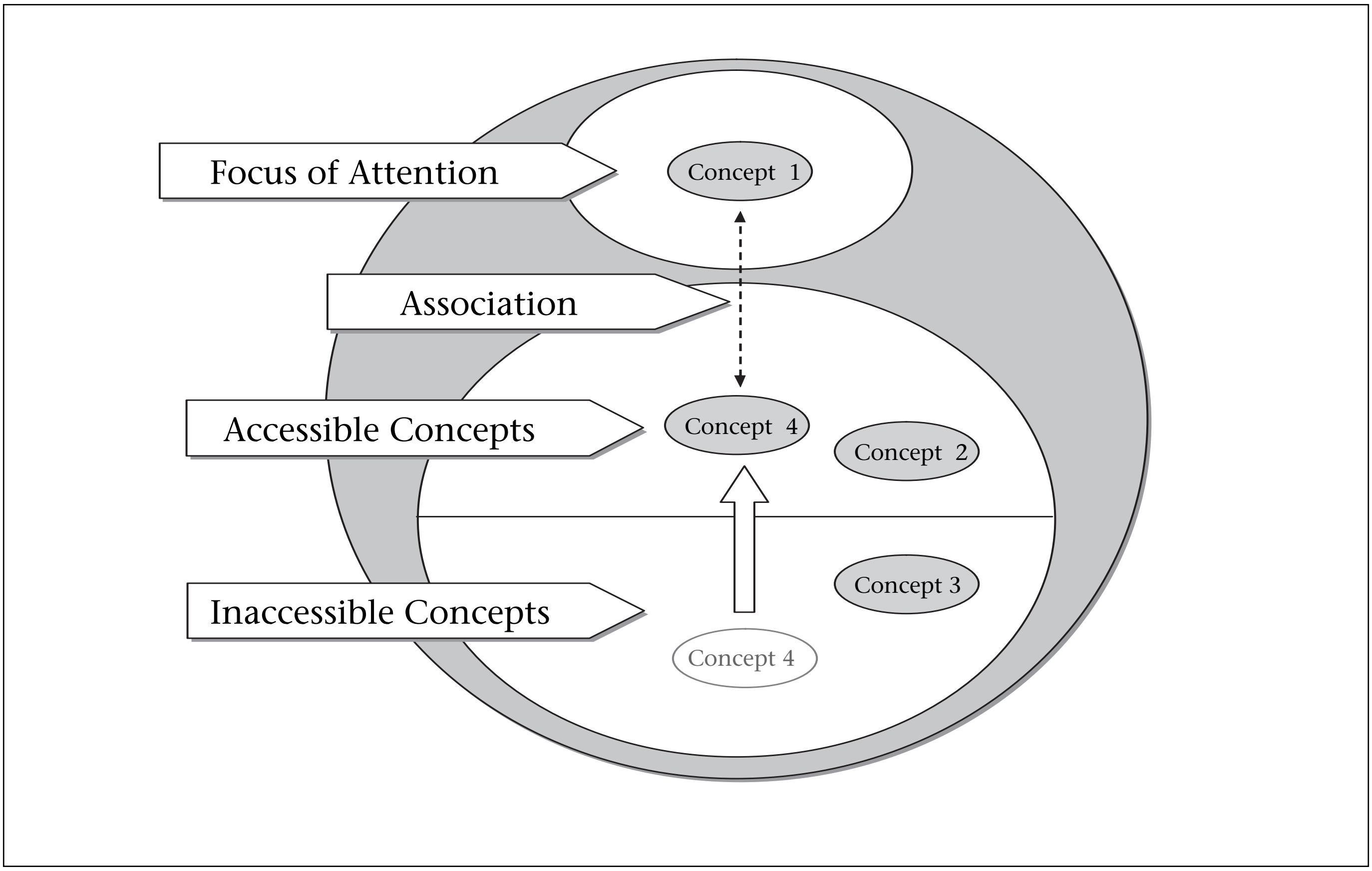

Figure 3. Accessibility by Association.

culus (Gordon and Hobbs 2003). Figure 4 is one of the more interesting examples of an axiom, the one that says that thinking of something raises the accessibility of associated concepts.

Figure 4 says that if concept c_1 is associated with concept c_2 for agent *p*, who has focus *f*, then if there is a change from c_1 not being in *p*'s focus *f* to c_1 being in *f*, then that causes an increase in the accessibility of c_2 for *p*. The use of logical expressions inside the predicates *cause* and *change* could be eliminated by reifying the *inm, accessibility*, and *change* relations, as in Hobbs (1985, 2003). This increase in accessibility may or may not be enough for *p* then to retrieve c_2.

It needs to be part of a theory of thinking or envisionment that agents can cause themselves to have a concept in their focus of attention. Then because of associations among concepts, agents have a causal structure they can manipulate to bring about retrievals from memory. This gives rise to strategies for remembering that involve calling to mind related concepts. For example, someone might place a box of dishwasher soap at the front door to remind him to turn on the dishwasher before leaving for work, relying on the natural association between dishwasher soap and dishwashers. We might try to remember someone's name by running through the letters of the alphabet and hoping that the first letter of the name is associated strongly enough with the name to cause the name to be retrieved.

A theory of goals would have to include an explication of a partial ordering of importance. A concept is more or less important to an agent at a particular time, depending on the relation of the concept to the agent's goals. The importance associates the concept to the goals.

There is at least a defeasible monotonic relation between the importance of a concept and its accessibility. The greater the importance of a concept, the more accessible it is. This relation is a key part of an explanation for why a man's forgetting his wedding anniversary might cause his wife to get angry. If it had been important to him, he would have remembered.

Ability, Trying, Succeeding, and Failing

A theory of goals and planning would also have to have an explication of the notions of ability, trying, succeeding, and failing. In other

$$
\begin{array}{l}
associated(c_1, c_2, p) \\
\supset (\exists f)[focus(f, p) \\
\quad \wedge\ (\forall e, t_1, t_2)(\exists a_1, a_2, t_3) \\
\qquad\quad [cause(change(\neg inm(c_1, f, t_1), inm(c_2, f, t_2), t_2), \\
\qquad\qquad change(a_1 = accessibility(c_2, p, t_2), a_2 = accessibility(c_2, p, t_3), t_3)) \\
\qquad\quad \wedge\ a_1 < a_2]]
\end{array}
$$

Figure 4. Axiom for Accessibility by Association.

work, we have at least sketched what the axiomatization would look like. All of these concepts concern an agent's manipulations of the causal structure of the world to achieve goals.

For an agent to *try* to achieve some eventuality is for the agent to perform actions that, it is believed, tend to cause the eventuality, where the eventuality is a goal of the agent. Furthermore, the agent should have the goal that those actions cause the eventuality. This rules out the case where the agent has a goal, performs actions that tend to cause the goal, but does not do these actions with the intention of bringing about the goal, but rather for some other reason.

To *succeed* in an attempt is to bring about that goal in actuality. To *fail* in an attempt is to attempt but not succeed.

Ability is much more difficult to characterize. We first begin with the notion of possibility. An eventuality is *possible* with respect to a set of constraints if the constraints do not entail the eventuality not occurring.

Planning is a matter of exploiting the causal structure of the world in order to achieve goals. These plans will typically require that, in addition to actions on the part of the agent, certain conditions that are beyond the control of the agent be true in the world. For example, we can drive a car to work if the streets aren't flooded. An agent is able to achieve some effect if and only if the required world conditions beyond the agent's control are right whenever the agent has the goal of achieving that effect. In other words, an agent is able to do something if that something is possible with respect to a set of constraints that includes the agent's desire to do it and the right world conditions being true.

Because people can cause concepts to be in focus, and this may cause them to remember other concepts, people have an ability to remember things. It is also possible for people to *try* to remember things, and thus to succeed or fail to remember things.

The Meaning of "Remember" and "Forget"

The English word *remember* can refer to a range of notions. At the simplest level, it can mean that the agent has the concept in memory and that it is accessible, but not necessarily in focus. In this sense, you remembered 20 minutes before encountering this sentence that Columbus discovered America in 1492. Even though the fact was not in focus, it was accessible in memory. Thus, to remember a concept is to have the concept in memory with accessibility above the memory threshold.

A somewhat stronger notion of remembering is when there has actually been a retrieval from memory. For a concept to be retrieved is for it to be remembered.

This rule was deliberately stated in the passive. There is no commitment to the agent's agency in the retrieval of a concept as we defined *retrieve*. Retrievals just happen. Thus, this notion of remembering covers cases where a fact simply pops into a person's head, with no prior effort or intention.

A stronger sense of "remember" is one in which the agent plays a causal role in the remembering. This happens when we are told to remember who the president of the United States is, and somehow immediately we do. This sense of "remember" is what is conveyed in imperatives like

> Remember that bears are unpredictable.

(Often such sentences are used to invite the hearer to draw an inference rather than retrieve something from memory, but in these cases it implies that it was already in memory.)

This definition of remembering is silent about whether the causality is immediate or there are intermediate actions on the agent's part designed to jog the memory. A stronger notion of remembering involves the latter. There is a distinct attempt to retrieve something from memory, and it succeeds. Since *succeed* as characterized above entails trying, we can simply say that to succeed in retrieving is to remember.

There are at least two levels of forgetting. In the simplest, the accessibility of a concept in memory has fallen below the memory threshold. To forget a concept is for the accessibility of the concept to change from being above the memory threshold to being below it.

It is a theorem in this theory that if an agent forgets something at a particular time, he does not remember it at that time, under any interpretation of remembering.

One might argue that another sense of "forget" occurs when something is not remembered at the appropriate time, even though it was accessible. For example, someone dashes into the surf, is pulled out to sea, is rescued, and says, "I forgot about the undertow." One could say the concept was accessible; it just wasn't accessed. But it is probably cleaner to say that its accessibility changed, since there will be many factors that induce changes in accessibility, and to stick with our first characterization of "forget."

Our characterizations of the meanings of "remember" and "forget" illustrate an important relation between core commonsense theories and the lexicon of English or any other language. The core theory of some knowledge area is constructed in a careful, coherent way, and the predicates axiomatized in the core theory can then be used to characterize the various uses of the related lexical items. This shows the way toward a deeper lexical semantics than has heretofore been possible.

Remembering to Do

Our plans for achieving goals spread across time. For example, the goal of eating dinner tonight might involve stopping at the grocery store on the way home. The timely performance of an action requires us to be consciously aware of the need to perform the action at the time of its performance. Since things cannot be retained continuously in the focus of attention, it is necessary to remember to do actions before doing them. Thus, as a precondition for doing an action, remembering to do it must also be a part of the plan of which the action is a part.

An action taking place at a particular time is enabled by being in the agent's focus of attention at that time. Thus, remembering to do something can become part of a plan, and hence an intention. As with all actions, a person can succeed or fail at remembering to do something.

Memory and Reasoning about Beliefs

A common and serious problem in classical AI theories of belief and reasoning about belief (for example, Moore 1985) is that of logical omniscience. The simplest ways of enabling an agent to draw conclusions from its beliefs have the side effect that an agent believes all the logical consequences of its beliefs. Thus, if someone knows the axioms of set theory, one knows all of mathematics.

This obviously does not correspond with our experience, and numerous researchers have devised theories that work around this unfortunate consequence in various ways (for example, Konolige 1985).

Our theory of memory provides a natural way to work around this. A classical theory of belief would say that if an agent believes P and believes that P implies Q, then the agent believes Q. In our theory of memory, we insist in addition that the beliefs be in the focus of attention. Thus, if an agent is focused on its belief P at time T and is at the same time focused on its believe that P implies Q, then the agent will believe Q immediately after t. Thus, inferences are limited by the premises that are in focus.

In fact, mathematicians often fail to prove theorems, even though they know all the required premises, because they do not see the premises as relevant to this problem and thus they do not come to be in focus.

Repressing

At least since the time of Freud, our commonsense theories include the notion that memories can be repressed. The passive formulation of *repressed* requires less in the way of ontology than the active action of *repressing*, so we will consider that first.

If a concept is repressed at a particular time for an agent, then the concept is in the agent's memory, but its accessibility is less than the memory threshold. Moreover, if a concept is repressed, it is unpleasant to the agent. (The predicate *unpleasant* would have to be explicated in a theory of emotions and a theory of goals.) Finally, the unpleasantness of the concept plays a causal role in the concept's being repressed.

It is problematic to say that an agent represses a memory. We may want to say in a theory of envisionment, or thinking, or consciousness, that agents are aware of what they are doing. But to store something in memory in a way that it cannot be accessed is as contradictory as being told not to think of an elephant. There are two ways around this problem. The first is to say that there are some actions that an agent may do without being conscious of them. The second, the Freudian approach, is to say that agents have within them subagents that can perform actions the superagent is not aware of. These two approaches are probably equivalent.

Discussion and Conclusions

In this article we have argued that the central challenge in commonsense knowledge representation research is to develop content theories that achieve a high degree of both *competency* and *coverage*. We described a new methodology for constructing formal theories in commonsense knowledge domains that complements traditional knowledge representation approaches by first addressing issues of coverage. We have shown how a close examination of a very general task (strategic planning) leads to a catalog of the concepts and facts that must be encoded for general commonsense reasoning. These can be sorted into a manageable number of coherent domains, one of which is the representational area of commonsense human memory. We can then elaborate on these concepts using textual corpus-analysis techniques, where the conceptual distinctions made in natural language are used to improve the definitions of the concepts that should be expressible in our formal theories. These representational areas can then be analyzed using more traditional knowledge representation techniques, as demonstrated in this article by our treatment of commonsense human memory.

Commonsense human memory is a particularly interesting domain with which to illustrate this approach. Although human memory has been extensively studied in psychology, there have been very few attempts at formal axiomatizations, and those that do exist (for example, Davis 1994) are extremely limited in scope. As in previous work, however, we also believe that the importance of inferential theories of commonsense memory will be most evident in future planning systems, where remembering is a crucial element in plans aimed at achieving goals over long periods of time.

Commonsense human memory is just one of the 30 commonsense psychology representational areas that we are formalizing in the context of this research effort. Significant challenges remain to be addressed, including issues surrounding the integration of all 30 of these representational areas within a larger reasoning framework. By taking an approach to knowledge representation that is based on an analysis of natural language, the opportunities ahead include the development of natural language processing systems that can more easily take advantage of commonsense theories and automated inference than is possible in current systems. The central vision, however, is one day to construct AI systems that know enough about the commonsense models that people have of themselves to better serve the needs of their users.

Acknowledgements

This article was developed in part with funds from the U.S. Army Research Institute for the Behavioral and Social Sciences under ARO contract number DAAD 19-99-D-0046, and in part with funds from the Advanced Research and Development Activity (ARDA). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Department of the Army or of ARDA.

References

Baron-Cohen, S. 2000. Theory of mind and autism: a fifteen year review. In *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, 2nd ed., ed. S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen. Oxford: Oxford University Press.

Clark, A. 1987. From Folk Psychology to Naive Psychology. *Cognitive Science* 11(2): 139–154.

Cohn, A.; and Hazarika, S. 2001. Qualitative Spatial Representation and Reasoning: An Overview. *Fundamenta Informaticae* 46(1–2): 2–32.

Collins, J. and Forbus, K. 1989. Building Qualitative Models of Thermodynamic Processes. Paper presented at the Third International Qualitative Reasoning Workshop, Stanford, Calif., May 1989.

Corcoran, R. 2001. Theory of Mind in Schizophrenia. In *Social Cognition in Schizophrenia,* ed. D. Penn and P. Corrigan. Washington, DC: American Psychological Association.

Davis, E. 1993. The Kinematics of Cutting Solid Objects. *Annals of Mathematics and Artificial Intelligence* 9(3–4): 253–305.

Davis, E. 1994. Knowledge Preconditions for Plans. *Journal of Logic and Computation* 4(5): 253–305.

Davis, E. 1998. The Naive Physics Perplex. *AI Magazine* 19(4): 51–79.

Goldman, A. 2000. Folk Psychology and Mental Concepts. *Protosociology* 14(2000): 4–25.

Gopnik, A.; and Meltzoff, A. 1997. *Words, Thoughts, and Theories.* Cambridge, Mass.: The MIT Press.

Gordon, A. 2001a. Strategies in Analogous Planning Cases. In *Proceedings, Twenty-Third Annual Conference of the Cognitive Science Society,* ed. J. Moore and K. Stenning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gordon, A. 2001b. The Representational Requirements of Strategic Planning. Paper presented at the

Fifth Symposium on Logical Formalizations of Commonsense Reasoning, New York, NY, 20–22 May.

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

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

Gordon, A.; Kazemzadeh, A.; Nair, A.; and Petrova, M. 2003. Recognizing Expressions of Commonsense Psychology in English Text. Paper presented at the Forty-First Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 7–12.

Gordon, Andrew S., and Hobbs, Jerry R. 2003. Coverage and competency in formal theories: A commonsense theory of memory. Proceedings, AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford, CA, March 2003.

Happe, F.; Brownell, H.; and Winner, E. 1998. The Getting of Wisdom: Theory of Mind in Old Age. *Developmental Psychology* 34 (2): 358–362.

Hayes, P. J. 1979. The Naive Physics Manifesto. In *Expert Systems in the Microelectronic Age,* ed. D. Michie. Edinburgh: Edinburgh University Press.

Hayes, P. J. 1984. The Second Naive Physics Manifesto. In *Formal Theories of the Commonsense World,* ed. J. Hobbs and R. Moore. Norwood, NJ: Ablex Publishers.

Hobbs, J. 1985. Ontological Promiscuity. In *Proceedings, Twenty-third Annual Meeting of the Association for Computational Linguistics, Chicago, Ill.,* 61–69. East Strasbourg, Penn.: Association for Computational Linguistics.

Hobbs, J. 2003. The Logical Notation: Ontological Promiscuity, Chapter 2 of *Discourse and Inference*. Available at http://www.isi.edu/~hobbs/disinf-tc.html.

Hobbs, J. 2002. Toward an Ontology of Time for the Semantic Web. Paper presented at the Workshop on Annotation Standards for Temporal Information in Natural Language, Third International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, 27 May.

Konolige, K. 1985. Belief and Incompleteness. In *Formal Theories of the Commonsense World,* 359–403, ed. J. Hobbs and R. Moore. Norwood, NJ: Ablex Publishing.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Moore, R. 1985. A Formal Theory of Knowledge and Action. In *Formal Theories of the Commonsense World,* 319–358, ed. J. Hobbs and R. Moore. Norwood, NJ: Ablex Publishing.

Nichols, S.; and Stich, S. 2002. How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In *Consciousness: New Philosophical Essays,* ed. Q. Smith and A. Jokic. Oxford: Oxford University Press.

Niles, I.; and Pease, A. 2001. Towards a Standard Upper Ontology. In *Proceedings of the Second International Conference on Formal Ontology in Information Systems (FOIS-2001),* ed. Chris Welty and Barry Smith. New York: Association for Computing Machinery.

Shanahan, M. 1995. A Circumscriptive Calculus of Events. *Artificial Intelligence Journal* 77(2): 249–284.

Silberztein, M. 1999a. Text Indexing with INTEX. *Computers and the Humanities* 33(3): 265–280.

Silberztein, M. 1999b. INTEX: A Finite State Transducer Toolbox. *Theoretical Computer Science* 231(1): 33–46.

van der Hoek, W.; and Wooldridge, M. 2003. Towards a Logic of Rational Agency. *Logic Journal of the IGPL* 11(2): 133–157.

Watt, S. 1995. A Brief Naive Psychology Manifesto. *Informatica* 19(4): 495–500.

Wellman, H. M.; and Lagattuta, K. H. 2000. Developing Understandings of Mind. In *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, 2nd ed., ed. S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen. Oxford: Oxford University Press.

Andrew S. Gordon is a research assistant professor at the University of Southern California Institute for Creative Technologies. His research interests include knowledge representation, natural language processing, and cognitive modeling. He is the author of the book *Strategy Representation: An Analysis of Planning Knowledge*. His email address is gordon@ict.usc.edu.

Jerry R. Hobbs is a research professor at the University of Southern California Information Sciences Institute. Formerly he was a principal scientist at SRI International. His research interests include encoding commonsense knowledge, knowledge representation, natural language processing, and discourse analysis. His email address is hobbs@isi.edu.