Toward a Large-scale Formal Theory of Commonsense Psychology for Metacognition

Jerry R. Hobbs

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
hobbs@isi.edu

Andrew S. Gordon

USC Institute for Creative Technolgies
13274 Fiji Way
Marina del Rey, CA 90292
gordon@ict.usc.edu

Abstract

Robust intelligent systems will require a capacity for metacognitive reasoning, where intelligent systems monitor and reflect on their own reasoning processes. A large-scale study of human strategic reasoning indicates that rich representational models of commonsense psychology are available to enable human metacognition. In this paper, we argue that large-scale formalizations of commonsense psychology enable metacognitive reasoning in intelligent systems. We describe our progress toward developing 30 integrated axiomatic theories of commonsense psychology, and discuss the central representational challenges that have arisen in this work to date.

Commonsense Psychology and Metacognitive Reasoning

Metacognition is defined in terms of commonsense psychological concepts. The term is meant to refer to the suite of human mental behaviors for monitoring and control, including the selection of goals, assessing progress toward those goals, adopting new strategies for achieving goals, and even abandoning a goal entirely. In providing examples of human metacognition in practice, situations where people employ reasoning strategies are often cited. For instance, metacognition is exhibited when a person sticks a note on their bathroom mirror so that they will remember to do something each morning, or when students assess the relative difficulty of exam questions in order to determine which ones to tackle first.

Gordon's large-scale analysis of human planning strategies (Gordon, 2004) revealed that all strategies have metacognitive aspects to them. In this work, 372 strategies that people use in their everyday lives to accomplish personal and professional goals were collected from ten different real-world planning domains (e.g. governance, warfare, and artistic performance). These strategies were then encoded as preformal definitions aimed at identifying the formal knowledge representations that will be required to enable the creation of cognitive models of human strategic reasoning. Of the 988 unique concepts that were required to author preformal definitions of these strategies, 635 were identified as pertaining to mental states and processes. Gordon (2002) organized these 635 concepts into 30 representational areas (e.g. Memory, Explanations, Expectations, Goal management, Plan adaptation, Execution control), a set which stands as the most comprehensive characterization of human commonsense psychological models available today. Figure 1 illustrates these 30 representational areas by clustering them around the central areas of Knowledge, Envisionment, Goals, Planning, and Execution.

There is a growing interest in trying to create intelligent systems that are themselves metacognitive, in that they monitor and control their own reasoning processes to respond proactively to problems and perform better with less need for human intervention. There may be several different metacognitive reasoning approaches that would be successful across different types of intelligent systems, depending on their computational tasks. However, intelligent systems designed to cooperate adaptively with humans will need to utilize representational models that can be aligned with those of their users. To build cooperative, adaptive intelligent systems that engage in metacognitive reasoning, it will be necessary to develop large-scale inferential models of human commonsense psychology.

This paper describes our efforts in authoring formal axiomatic theories of human commonsense psychology. First, we describe how we elaborated the representational requirements for strategic reasoning identified by Gordon (2004) by conducting a large-scale analysis of English phrases related to mental states and processes. Second, we discuss our progress in authoring axiomatic theories for the 30 representational areas and the lessons learned so far. We then conclude with directions for future work.

Authoring Large-Scale Theories of Commonsense Psychology

Our authoring approach was to identify clearly the representational requirements of large-scale theories of commonsense psychology using analytic and empirical techniques before beginning to encode this knowledge as formalisms in first-order predicate calculus. We began with

Figure 1. The Thirty Representational Areas of Commonsense Psychology

an analysis of Gordon's 30 representational areas of commonsense psychology, and the 635 concepts that were sorted among them. Because these concept lists contained only the terms necessary for the adequate definition of the real-world strategies that were analyzed as part of that study, some redundancy and gaps were evident. For example, the representational area of Managing Expectations (dealing with the events that people expect to happen in future states) listed the term *Expectation violation*, referring to the mental event of being surprised by something that occurs, but does not include in the list of eight terms a corresponding concept for *Expectation confirmation*, referring to the mental event of realizing that one's expectations have been met.

To elaborate these 635 concepts as well as reduce the redundancy that was apparent, we decided to use natural language as additional empirical evidence for the commonsense psychological concepts that were necessary to manipulate in formal theories. Beginning in Summer 2002 and ending in Fall 2004, we conducted a large-scale effort to identify every possible way to express in the English language all of the concepts related to the 30 representational areas. This work was completed by first identifying multiple ways to express each of the 635 concepts in the English language. These examples were then used as a launching point for large group discussions aimed at eliciting even more examples (typically dozens for each concept). The resulting sets of examples were then organized to determine cases where the existing concept set for a representational area lacked a concept that was expressible in language (a term needed to be added) and cases where language made no distinction between two existing concepts (two concepts needed to be combined into one). Computational linguistics graduate students then identified full sets of synonymous expressions for each of the examples, and authored finite-state transducers capable of automatically tagging expressions of commonsense psychology concepts in English text. Gordon et al. (2003) evaluated the quality of this approach to authoring finite-state transducers for four of the 30 representational areas, reporting an average precision performance of 95% and an average recall performance of 82%.

Through this approach, a final set of 528 concepts was identified that describe the representational requirements of formal theories of commonsense psychology. With these requirements in hand, we then began the process of applying more traditional knowledge representation methods of formalization and axiomization.

Formal Theories

We have formalized fourteen of the knowledge domains so far; here we focus on the first nine we completed: Memory, Knowledge Management, Envisionment, Goals,

Goal Themes, Plans, Plan Elements, Scheduling, and Execution Monitoring. The first two provide a more extensive model of a belief system than has been traditional in theories of belief. The third is a first cut at an axiomatization of what it is to think about something. The fourth, fifth, sixth and seventh begin to address notions of intentionality, and the last two deal with the interaction of intentions with reality. Thus, while we discuss only nine out of thirty commonsense domains here, they represent a broad range of the most basic features of human cognition, as we normally conceive of it in everyday life.

A traditional problem in axiomatizing a domain is the problem of determining exactly what the coverage should be. The list of 528 concepts derived from the study of strategies, together with their meanings, provides the answer to this question. The formal theories must identify the key central underlying concepts and define the other concepts in terms of these. A further constraint on the construction of formal theories comes from the very size of the set of concepts to be explicated. There are logical relations among the representational areas, e.g. reasoning about Goal Management requires an understanding Goals. Thus, more basic theories must explicate the concepts required by the theories that depend on them. The entire effort must yield a coherent set of theories. We cannot take shortcuts early in the effort without undercutting our coverage later.

We should emphasize that our aim is to produce logical theories of commonsense psychological concepts, as outlined by strategies and language use, rather than to formulate predictive psychological models (e.g. a commonsense model of human memory, rather than a psychological model of human memory performance).

In what follows we describe the high points of the theories, indicating the coverage and the key ideas.

Memory

In most theories of belief, beliefs are not distinguished as to their availability to inference processes (e.g., Moore, 1985). But this is, sadly, an idealization that falls short of human experience. We are always forgetting things it would be in our interests to remember, remembering things just in time or a little too late, being reminded of things out of the blue, and trying to recall things when we need them.

The first step we take beyond the traditional theories is to introduce an internal structure for the "mind" that contains the beliefs. A person's mind has a "focus of attention," or simply "focus", and a "memory". Concepts are in one or the other. In our theories of other areas, such as Envisionment, certain actions on concepts require the concepts to be in focus. For example, you can't consider the consequences of some event without attending to the event consciously. Concepts can be stored in memory and can be retrieved from memory.

The second step we take beyond the traditional theories is to introduce a notion of the "accessibility" of concepts in memory. Accessibility is a partial ordering. There is a memory threshold such that when the accessibility of a concept falls below it, it can no longer be retrieved. The concept has been forgotten. The greater the accessibility, the easier it is to retrieve into focus.

We posit a general notion of "association" between concepts that encompasses relations like implication, among others. When a person retrieves a concept, this action increases the accessibility of the concepts with which it is associated. This gives the person some control over the retrieval of forgotten concepts. With this vocabulary, we are now in a position to state in logic the strategies people have for memorization and to verify formally that they are effective strategies.

A more complete account of our Theory of Memory is presented in Gordon and Hobbs (2004).

Knowledge Management

Our theory of Knowledge Management concerns the properties of beliefs and how they are organized. First, we have a standard theory of belief. Beliefs are relations between an agent and a proposition. Agents can use modus ponens; that is, there is a defeasible inference that if an agent believes P and believes P implies Q, then the agent believes Q. This is only defeasible since we would have logical omniscience otherwise. More particularly, if an agent believes P and has P in focus, and believes P implies Q (in focus or not), then defeasibly the agent will come to have Q in focus and will believe it.

Other central properties of belief await the development of other theories. The fact that people generally believe what they perceive awaits a Theory of Perception. The fact that they often believe what they are told should be handled in a Theory of Communication. The fact that we act in a way that tends to optimize the satisfaction of our desires given our beliefs will be dealt with as we develop our Theory of Plans more extensively.

Propositions can be proved from other propositions. Partial proofs consisting of plausible propositions tend to justify beliefs. Knowledge is, or at least entails, justified true belief (Chisolm, 1957, although cf. Gettier, 1963).

In addition, we have sketched out a theory of "graded belief". Agents can believe propositions to some degree. Degree of belief is a partial ordering. Some of the key properties of graded belief are that degree of belief does not diminish under modus ponens, that the degrees of belief of a proposition and its negation vary inversely, and that multiple independent supports for a proposition tend to increase its degree of belief. Graded beliefs over some threshold become full-fledged beliefs.

We have axiomatized graded belief in a very abstract and noncommittal manner, in a way that, for example, accommodates Friedman and Halpern's (1999) approach to nonmonotonic reasoning.

We define a notion of "positive epistemic modality". Positive epistemic modalities are preserved under modus ponens. In addition to belief and graded belief, suspecting (believing P more than ~P) and assuming are positive epistemic modalities.

We define a "sentence" to be a set of propositions, a subset of which constitutes the "claim" of the sentence. For example, in the sentence "A man works", the propositional content is that there is an x such that x is a man and x works. Only the latter proposition is the claim of the sentence.

The notion of "knowledge domain" is a difficult one to explicate, but at a first cut we characterize a knowledge domain by a set of predicates. For example, the domain of baseball would be defined by a set of predicates including "batter", "pitches", and so on. A knowledge domain is then a collection of sentences whose claims have predicates in that set. A fact is a sentence whose propositions are true.

The notion of expertise in a domain is also very complex to define, but we can postulate some properties of a "more expert than" relation, e.g., one agent who knows all the facts in a domain that another agent knows is more expert in that domain.

Mutual belief can be defined in the standard way. If a community mutually believes that P, then each member believes that P, and the community mutually believes that it mutually believes that P. Mutual belief can be extended to sentences and thus to knowledge domains. Then we can talk about communities and the knowledge domains they share. One of the most important kinds of knowledge we have is knowledge about who knows what, and much of this is inferred from our knowledge of the communities the agents belong to. For example, we believe that American citizens know the basic facts about the American government, and we believe that AI researchers know about the frame problem.

World Envisionment

The cognitive process of "thinking" is very hard indeed to pin down formally. We can (and do) define "thinking of" a concept as having that concept in focus. We can and do define "thinking that" a proposition is true, as in "John thinks that the world is flat", as having a proposition in focus that one believes. But "thinking about" a concept, an entity, or a situation can cover a broad range of complex cognitive processing. Nevertheless, we can begin to pin down one variety of such cognitive processing – envisioning, or beginning with a situation and working forwards or backwards along causal chains for the purposes of prediction or explanation.

We base our treatment of envisionment on the formalization of causality in Hobbs (*in press*). This paper introduces the notion of a "causal complex" for an effect; essentially, if everything in the causal complex holds or happens, then the effect happens, and for any event or state in the causal complex there is some situation in which toggling it changes the effect. We then say one eventuality is "causally involved" with an effect if it is in a causal complex for the effect. Two eventualities are "causally linked" if there is a chain of "causally involved" relations from one to the other. A "causal system" consists of a set of eventualities and a set of "causally involved" relations among them. A causal system is connected if every two eventualities in it are causally linked.

An agent has an "envisioned causal system slice" when the agent is thinking of all the eventualities in a connected causal system, where the agent either believes or is thinking of the causal relations among them. Two envisioned causal system slices are contiguous if one results from the other by adding or deleting a causally-involved relation. When the eventuality that is added is a cause, the agent is doing explanation; when it is an effect, the agent is doing prediction.

An "envisioned causal system" is then a temporal sequence of envisioned causal system slices. It is a kind of movie of what the agent is thinking of as the agent reasons forwards and backwards along causal chains.

Envisioned causal systems can be purely fictional or imaginary – what will I do if I win the lottery? But an especially important subclass of envisioned causal systems begins with the world as perceived where the causal relations are all believed to be true. In this case, each envisioned causal system slice is the agent's "current world understanding".

The Theory of Envisionment thus provides the formal vocabulary for us to talk about the agent's instrumental cogitation in trying to figure out what's going on in the world now, why, and what will happen next.

Goals and Goal Themes

Cognition meets action in the Theories of Goals and Planning. People are intentional agents. That is, they have goals and they devise, execute, and revise plans to achieve these goals. Other computational agents besides people can be viewed as planning agents as well, including complex artifacts and organizations.

Agents use their knowledge about causation and enablement to construct plans. In the traditional AI picture of planning (Fikes and Nilsson, 1971), agents decompose goals into subgoals by determining everything that enables the goal (the prerequisites) and positing these as subgoals, and by finding some complex of actions and other eventualities that will cause the goal to occur (the body). The resulting structure is a "plan" to achieve the goal. More precisely, if an agent has a goal and the agent believes some eventuality enables that goal, then the agent will adopt that eventuality as a goal as well. If an agent has a goal and the agent believes some action will cause the goal to be satisfied once it is enabled, then defeasibly the agent will adopt the action as a goal – "defeasibly" because there may be more than one way to achieve the goal. Moreover, having the goal causes the adoption of the subgoal.

It is sometimes said that it is a mystery where goals come from. But it is easy to get around this difficulty by stipulating that agents have "thriving" as their top-level goal. All other goals can then be generated via beliefs about what will cause the agent to thrive. In the case of most people, this will involve surviving, but it is certainly possible for people and other agents to have the belief that

they best thrive when the group they belong to thrives, thereby placing other goals above surviving. Thriving does not necessarily imply surviving.

Individual persons and individual computational agents are not the only kinds of computational agents. It is also possible for collectivities of agents to have goals and to develop plans for achieving these goals. For example, the organization General Motors can be viewed as an intentional agent whose goal is to sell cars. Its plan involves manufacturing and marketing cars. These plans must bottom out in the actions of individual persons or devices, or in states or events that will happen at the appropriate time anyway. The structure of an organization frequently reflects the structure of the plan the organization implements. For example, a car company might have a manufacturing and a marketing division.

We can define a collection of agents having a shared goal in terms of each of the members having the corresponding individual goal and there being mutual knowledge among the members that the collection as a whole has that goal.

We define the notion of a "goal theme" (Schank and Abelson, 1977) as a set of goals that a set of agents has. Goal themes are useful for predicting other agents' goals from minimal knowledge about them, namely, the groups they belong to. If you see an enemy soldier, you know that he has the goal of killing you. The set of agents can be characterized in many ways. Goal themes can result from an agent's nationality, from a role in an organization, from a relationship, or from a lifestyle choice, for example.

Plans, Plan Elements, and Scheduling

Plans are what turn beliefs into actions. An agent figures out what to do to achieve the goals, and then does it. But plans go through a number of stages from conception to execution, and if we are going to be able to make as subtle distinctions as people make about that process, we will have to explicate these different degrees of commitment.

We begin with a simplified model that distinguishes between the belief system and the plan. Agents reason about actions that would result in their goals being satisfied; this is a matter of reasoning about their beliefs, resulting in beliefs in large-scale causal rules, or "plans in waiting". At some point, an agent "commits" to some of these plans in waiting, and they become "plans in action". This act of committing to a plan we call "deciding to". The agent is continually deciding to perform certain actions by committing to certain goals and to certain plans for achieving the goals.

In commonsense reasoning it is possible for agents to directly cause events. For example, when a dog gets up and crosses a yard, we might say that there were certain events in the dog's brain that cause it to cross the yard. But more often we simply think of the dog itself as initiating that causal chain. So we introduce the notion of "directly cause" as a place where planning can bottom out. Events that are directly caused by an agent are like the executable actions in planning systems; they are the actions an agent can just do.

Desires and preferences can be modeled as beliefs about the efficacy of certain states and events causing the agent to thrive, or to achieve some lower-level goal. Like other causal beliefs, they play a role in the plans the agent derives for achieving goals, and thus often find their way into the plans that are actually executed.

Note that this view of action as the execution of plans goes beyond our everyday notion of planned behavior – when someone whose head itches scratches it, this does not seem like a matter of planning. But it certainly is instrumental behavior that taps into the underlying causal structure of the world, and as such can be represented at a formal level as the execution of a plan.

Planning is a kind of envisioning where the initial envisioned causal system is simply the goal and the successive envisioned causal systems are hierarchical decompositions of the goal, until the agent reaches causes that are directly caused actions of the agent's.

Scheduling is a matter of adding a consistent set of temporal parameters to a plan. The constraints on scheduling derive from various inabilities to perform several kinds of actions at the same time, and must be specified in domain theories. Once we have specified a schedule for an agent's plan, we can talk about the agent's "schedule capacity" and "next unscheduled moment". A great deal of discourse about scheduling involves deadlines for tasks and the "slack" one has for completing a task – "we have plenty of time", "he finished in the nick of time". We define slack as the time between the believed or estimated completion of a task and the deadline for that task.

Execution Control

Once we begin to execute a plan, we monitor the environment to see if the goals are being achieved. If they are not or if other events intervene, we modify, postpone, suspend, abort, resume, restart, or do a number of other actions. The Theory of Execution Control is about the agent's manipulations of the plan as it unfolds in time. In linguistics, these notions go under the name "aspect" – is the action completed, continuing, just started, and so on. Perhaps the best treatment of aspect from an AI perspective is that of Narayanan (1999), who develops a detailed model of processes in terms of Petri nets, identifying which parts of processes each aspect describes. In our model, we carry this to one more level of detail by defining various aspects in terms of hierarchical plans.

Some basic concepts have to be defined before we can address the central issue. The "left fringe" of a plan is the set of actions that can be initiated immediately, before any other actions in the plan are executed. The "remaining plan" at any given instant is that part of the plan that has not yet been executed. We also define the notion of two specific plans being instantiations of the same abstract plan, perhaps just displaced in time.

We can then say that to "start" a plan is to execute an action in its left fringe. To "stop" the execution of a plan is to change from executing it to not executing it. To "continue" a plan at a given time is to execute an action in the left fringe of the remaining plan. To "resume" a plan is to start a temporally displaced instantiation of the remaining plan of the same abstract plan after it has been stopped. To "restart" a plan is to "start" an instantiation of the same abstract plan from the beginning. To "pause" in a plan is to stop and then to resume. A plan is "ongoing" if it is either being executed or there is a pause in its execution. To "suspend" a plan is to stop it with the intention of resuming it. To "abort" a plan is to stop it with the intention of not resuming it. To "complete" the execution of a plan is to have executed every action in it.

We have also begun to explicate what it is to follow "instructions", a document, broadly construed, whose content is an abstract plan, and to put on a "performance", or the execution of a plan for the display to others.

Future Work

Having completed the first fourteen of the thirty component theories of commonsense psychology, our first priority for future work is to complete the remaining sixteen theories. As with the first nine, we anticipate that many of the remaining theories will challenge many of the simplifying assumptions that have traditionally been made in formal knowledge representation research. For other areas where little previous formalization work exists, we hope that exploring these areas will create an interest in the development of new competing theories, and hopefully a renewed interest in authoring formal content theories of commonsense reasoning within the field in general.

Our second priority for future work is the validation of coverage and competency of these component theories for reasoning about human metacognition. We are particularly interested in validating these theories by using them to derive formal proofs of human metacognitive strategies (Swanson & Gordon, *this volume*). Our aim is to demonstrate that this large-scale formal theory of commonsense psychology closely parallels the knowledge that is employed by people when making judgments about the appropriateness of any given metacognitive strategy for achieving reasoning goals.

Our third priority for future work is to develop practical intelligent systems that utilize a large-scale formal theory of commonsense psychology to engage in metacognitive reasoning in service of their users' goals. In particular, we are interested in applications that are able to capitalize on the substantial natural-language resources that were created as a product of our authoring methodology. The finite-state transducers that were created as part of this work can be reliably used to draw correspondences between English expressions of commonsense psychological concepts and their formal definitions, and hold the promise of enabling a new breed of natural language processing systems that effectively integrate automated commonsense reasoning with empirical methods.

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

Chisholm, R. (1957) Perceiving: A Philosophical Study, Ithaca, NY: Cornell University Press.

Fikes, R., & Nilsson, N. (1971) STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence* 2:189-208.

Friedman, N., & Halpern, J. (1999) Plausibility Measures and Default Reasoning: An Overview *Proceedings 14th Symposium on Logic in Computer Science*, Trento, Italy, July 1999.

Gettier, E. (1963) Is Justified True Belief Knowledge? *Analysis* 23: 121-123.

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

Gordon, A. (2002) The Theory of Mind in Strategy Representations. *24th Annual Meeting of the Cognitive Science Society* (CogSci-2002). Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, A. & Hobbs, J. (2004) Formalizations of Commonsense Psychology. *AI Magazine* 25: 49-62.

Gordon, A., Kazemzadeh, A., Nair, A., and Petrova, M. (2003) Recognizing Expressions of Commonsense Psychology in English Text. *41st Annual Meeting of the Association for Computational Linguistics* (ACL-2003) Sapporo, Japan, July 7-12, 2003.

Hobbs, J.. (*in press*) Toward a Useful Concept of Causality for Lexical Semantics. To appear in *Journal of Semantics*.

Moore, R. (1985) A Formal Theory of Knowledge and Action, in J. Hobbs and R. Moore, (eds.) *Formal Theories of the Commonsense World*, pp. 319-358, Ablex Publishing Corp., Norwood, New Jersey.

Narayanan, S. (1999) Reasoning About Actions in Narrative Understanding, *International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann.

Schank, R., Abelson, R. (1977) *Scripts, Plans, Goals, and Understanding*. Mahwah, NJ: Lawrence Erlbaum Associates.

Swanson, R. & Gordon, A. (2005) Automated Commonsense Reasoning About Human Memory. 2005 AAAI Spring Symposium on Metacognitive Computing. March 21-23, Stanford, CA.