Encoding Knowledge of Commonsense Psychology

Jerry R. Hobbs **Andrew S. Gordon**
Information Sciences Institute Institute for Creative Technologies
University of Southern California
Marina del Rey, California 90292
hobbs@isi.edu gordon@ict.usc.edu

Abstract

An analysis of human planning strategies reveals that much of the knowledge that underlies intelligent planning involves commonsense psychology, the way that people think that they think. In this paper we describe our continuing effort to formalize a large-scale theory of commonsense psychology as 30 interrelated content theories in first-order logic. This paper discusses key aspects of the 16 content theories that we have completed, focusing on those that provide an account of how knowledge and intention lead to action, namely, memory, knowledge management, envisionment, goals, planning, and execution. Some of these areas present challenges to many of the simplifying assumptions that have traditionally been made in formal knowledge representation research; others are areas of commonsense knowledge where few formal treatments have previously been attempted.

1 Introduction

In previous papers (Gordon and Hobbs, 2003, 2004) we have described a methodology for determining what knowledge should be included in the knowledge base for an intelligent agent, capable of constructing and executing plans to achieve its goals. An intelligent agent is at least a planning mechanism, so Gordon (2004) asked what concepts are necessary for the common strategies that people use in achieving their goals. He investigated ten different domains, including politics, personal relationships, artistic performance, and warfare, and collected 372 strategies. He authored representations of these strategies in order to identify a controlled vocabulary involving of concepts. These concepts were categorized into 48 different representational areas, such as sets, space, and time. Thirty of the representational areas, involving 635 concepts, were concerned with commonsense psychology; among these are memory, knowledge management, planning, and so on. This result by itself demonstrates the very great importance of commonsense psychology in the construction of intelligent agents.

Gordon et al. (2003) then, to define further each of the representational areas, augmented the list of concepts by investigating the English language expressions for concepts in each area. The result was a list of 528 concepts, a set that identifies the target coverage of a formal theory of commonsense psychology.

The authors began the development of formal theories that would encompass this list of concepts. In our earlier work (Gordon and Hobbs, 2003), we described the first theory we constructed, memory, as an illustration of the method. We have now completed 16 of the 30 theories, and this paper provides an overview of this work at the halfway mark.

2 Formal Theories

We have formalized sixteen of the knowledge domains so far. Here we focus on those that provide an account of how knowledge and intention lead to action: Memory, Knowledge Management, Envisionment, Goals, Planning, and Execution. 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 and fifth begin to address notions of intentionality, and the last one deals with the interaction of intentions with reality. Thus, while we discuss only some 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 and language use, together with their meanings, provides the answer to this question. The formal theories must identify the key 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., the area of Goal Themes depends on 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 formalizations of commonsense psychological theories that encode the way that people that they think. These theories may differ markedly from current scientific theories of cognitive processes, which are themselves sometimes formalized as computational process models. By focusing on commonsense (rather than scientific) knowledge, our aim is descriptive and predictive of the inferences that people make about mental states and processes, without making a strong commitment as to the underlying cognitive mechanisms of explanation and prediction in which these commonsense theories are employed. For example, in authoring a commonsense theory of human memory, our intent is to encode the inferences that people use to make explanations and predictions about the behavior of human memory, not to formalize the processes of human memory in a manner that is competitive with current scientific theory.

3 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 that are "in mind" (*inm*) are in focus or in memory. 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.

$$
\begin{aligned}
&person(p) \supset (\exists\, x) mind(x,p)^{1} \\
&mind(x,p) \supset (\exists\, f)[focus(f,p) \wedge part(f,x)] \\
&mind(x,p) \\
&\quad \supset (\exists\, m)[memory(m,p) \wedge part(m,x)]
\end{aligned}
$$

To think of something is to have it in focus.

$$think\text{-}of(p,c) \equiv (\exists\, f)[inm(c,f) \wedge focus(f,p)]$$

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.

A more complete account of our Theory of Memory is presented in Gordon and Hobbs (2003, 2004).

4 Knowledge Management

This theory concerns the properties of beliefs and how they are organized. We first of all 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 will 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), assuming, and mutual belief are all positive epistemic modalities.

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. We can talk about communities and their mutual beliefs. One of the most im-

[1] Axioms without outermost quantifiers are assumed to be universally quantified over the variables occurring in the antecedent of their highest level implication.

portant 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.

5 Envisionment

The cognitive process of "thinking" is very hard indeed to pin down formally. Above, we defined "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 believing the proposition. 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 (2001, 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.

We define a causal system as a set of eventualities (*evs*) together with a set of "causally involved" relations (*rels*) among them.

$$\begin{array}{l}
\mathit{causal\text{-}system}(s) \\
\supset\ (\exists\, s_1, s_2)[\mathit{evs}(s) = s_1\ \wedge\ \mathit{rels}(s) = s_2] \\
{}[\mathit{evs}(s) = s_1\ \wedge\ \mathit{rels}(s) = s_2] \\
\equiv\ [\mathit{causal\text{-}system}(s) \\
\quad \wedge\, (\forall\, e)[\mathit{member}(e, s_1) \\
\quad\quad \supset\ [\mathit{eventuality}(e) \\
\quad\quad\quad \wedge\, (\exists\, e_1, r)[\mathit{member}(r, s_2) \\
\quad\quad\quad\quad \wedge\, [\mathit{causally\text{-}involved}'(r, e_1, e) \\
\quad\quad\quad\quad\quad \vee\, \mathit{causally\text{-}involved}'(r, e, e_1)]]]] \\
\quad \wedge\, (\forall\, r)[\mathit{member}(r, s_2) \\
\quad\quad \supset\ (\exists\, e_1, e_2)[\mathit{member}(e_1, s_1) \\
\quad\quad\quad \wedge\, \mathit{member}(e_2, s_1) \\
\quad\quad\quad \wedge\, \mathit{causally\text{-}involved}'(r, e_1, e_2)]]]
\end{array}$$ [2]

That is, the eventualities in a causal system all participate in some *causally-involved* relation in the causal system, and every *causally-involved* relation in the causal system is between two eventualities in the causal system.

Two eventualities e_1 and e_2 are "causally linked" in a set of "causally involved" relations if there is a chain of relations in s between e_1 and e_2, regardless of direction.

$$\begin{array}{l}
\mathit{causally\text{-}linked}(e_1, e_2, s) \\
\equiv\ [(\exists\, r)[\mathit{causally\text{-}involved}'(r, e_1, e_2) \\
\quad\quad \wedge\, \mathit{member}(r, s)] \\
\quad \vee\, (\exists\, r)[\mathit{causally\text{-}involved}'(r, e_2, e_1) \\
\quad\quad \wedge\, \mathit{member}(r, s)] \\
\quad \vee\, (\exists\, e_3, r)[[\mathit{causally\text{-}involved}'(r, e_1, e_3) \\
\quad\quad\quad\quad \vee\, \mathit{causally\text{-}involved}'(r, e_3, e_1)] \\
\quad\quad \wedge\, \mathit{member}(r, s) \\
\quad\quad \wedge\, \mathit{causally\text{-}linked}(e_3, e_2, s - \{r\}]]
\end{array}$$

A "connected causal system" is a causal system in which all eventualities are causally linked by relations in the causal system.

$$\begin{array}{l}
\mathit{connected\text{-}causal\text{-}system}(s) \\
\equiv\ [\mathit{causal\text{-}system}(s) \\
\quad \wedge\, (\forall\, e_1, e_2)[[\mathit{member}(e_1, \mathit{evs}(s)) \\
\quad\quad\quad \wedge\, \mathit{member}(e_2, \mathit{evs}(s))] \\
\quad\quad \supset\ \mathit{causally\text{-}linked}(e_1, e_2, \mathit{rels}(s))]]
\end{array}$$

Multiple effects of a cause can be in the same connected causal system, whereas alternative effects of a cause may not be. The same is true for multiple causes of an effect.

An envisioned causal system is a causal system that some agent thinks of and extends forward and backward by prediction and explanation and by exploring alternatives. We first define an *ecs-slice*, that is, a causal system as it is envisioned at any given instant.

$$\begin{array}{l}
\mathit{ecs\text{-}slice}(s, a, t) \\
\equiv\ [\mathit{causal\text{-}system}(s) \\
\quad \wedge\, (\forall\, e)[\mathit{member}(e, \mathit{evs}(s)) \\
\quad\quad\quad \supset\ \mathit{think\text{-}of}(a, e, t)] \\
\quad \wedge\, (\forall\, r)[\mathit{member}(r, \mathit{rels}(r)) \\
\quad\quad\quad \supset\ [\mathit{believe}(a, r)\ \vee\ \mathit{think\text{-}of}(a, r)]]]
\end{array}$$

The agent a is (consciously) thinking of every eventuality in the envisioned causal system at time t. a may or may not be thinking of all the causal relations. They may be merely believed, not even necessarily accessible. For example, when we rapidly assess a situation by intuition, we are using causal knowledge we may not even be aware of. On the other hand, a may consciously be thinking of and considering the consequences of a causal rule he does not believe. In our full treatment, we also introduce a set of background assumptions that are not contradicted during the process of envisioning.

Next we characterize various ways in which two causal system can be contiguous.

$$\begin{array}{l}
\mathit{causally\text{-}involved}'(r, e_1, e) \\
\quad \wedge\, \mathit{member}(e, \mathit{evs}(s_1)) \\
\quad \wedge\, \neg \mathit{member}(e_1, \mathit{evs}(s_1)) \\
\quad \wedge\, \mathit{evs}(s_2) = \mathit{evs}(s_1) \cup \{e_1\} \\
\quad \wedge\, \mathit{rels}(s_2) = \mathit{rels}(s_1) \cup \{r\} \\
\supset\ \mathit{contig\text{-}cs}(s_1, s_2)
\end{array}$$

That is, adding an explanation of an element of a causal system results in a contiguous causal system.

$$\begin{array}{l}
\mathit{causally\text{-}involved}'(r, e, e_1) \\
\quad \wedge\, \mathit{member}(e, \mathit{evs}(s_1)) \\
\quad \wedge\, \neg \mathit{member}(e_1, \mathit{evs}(s_1)) \\
\quad \wedge\, \mathit{evs}(s_2) = \mathit{evs}(s_1) \cup \{e_1\}
\end{array}$$

[2] A primed predicate (e.g., *causally-involved*′(r, e, e_1)) says that its first argument (r) is a reification of the eventuality of its unprimed predicate (*causally-involved*) being true of its other arguments (e, e_1).

$$
\begin{array}{l}
\quad \wedge\, rels(s_2) = rels(s_1) \cup \{r\} \\
\supset\ contig\text{-}cs(s_1, s_2)
\end{array}
$$

That is, adding a prediction from an element of a causal system results in a contiguous causal system.

$$
\begin{array}{l}
causally\text{-}involved'(r_1, e_1, e) \\
\quad \wedge\, causally\text{-}involved'(r_2, e_2, e) \\
\quad \wedge\, member(e, evs(s_1)) \wedge member(e_1, evs(s_1)) \\
\quad \wedge\, \neg member(e_2, evs(s_1)) \\
\quad \wedge\, member(r_1, rels(s_1)) \\
\quad \wedge\, \neg member(r_2, rels(s_1)) \\
\quad \wedge\, evs(s_2) = [evs(s_1) - \{e_1\}] \cup \{e_2\} \\
\quad \wedge\, rels(s_2) = [rels(s_1) - \{r_1\}] \cup \{r_2\} \\
\supset\ contig\text{-}cs(s_1, s_2)
\end{array}
$$

That is, substituting one cause for another in a causal system results in a contiguous causal system. Similarly, substituting one effect for another in a causal system results in a contiguous causal system. These two rules give us a way of exploring alternatives in causal systems. We can think of e in the last axiom as a branch point. There can be different likelihoods associated with the different alternatives, including a likelihood of zero. Some branches are more likely than others.

The *contig-cs* relation is symmetric.

$$contig\text{-}cs(s_1, s_2)\ \equiv\ contig\text{-}cs(s_2, s_1)$$

Thus, we can proceed by deleting causes and effects as well as adding them.

These are not necessarily the only ways in which causal systems can be contiguous, but they are the principal ones.

Now we can define an envisioned causal system (*ecs*) as a sequence of envisioned causal system slices where the causal systems are contiguous and where there is a change of state from each causal system slice being envisioned to the next one in the sequence being envisioned.

$$
\begin{array}{l}
ecs(c, a, T) \\
\equiv\ (\forall\, s)[member(s, c) \\
\qquad \supset\ ecs\text{-}slice(s, a, t)\ \wedge\ in(t, T)] \\
\quad \wedge\, (\exists\, f)[map(f, ints(1, |c|), c) \\
\qquad \wedge\, (\forall\, i)[member(i, ints(1, |c|)) \\
\qquad\quad \supset\ [i = |c| \\
\qquad\qquad \vee\, [contig\text{-}cs(f(i), f(i+1)) \\
\qquad\qquad\quad \wedge\, (\exists\, t_1, t_2)[change(ecs\text{-}slice(f(i), a, t_1), \\
\qquad\qquad\qquad\qquad ecs\text{-}slice(f(i{+}1), a, t_2))]]]]]
\end{array}
$$

Here, c is a set of ecs-slices. $ints(1, |c|)$ is the sequence of integers from one to the size of c. $map(f, ints(1, |c|), c)$ says that f is a function mapping this sequence into the set c. We assume the ontology of time in Hobbs and Pan (2004).

An "envisioned causal system" is thus 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. It is also a way for an agent to consider hypotheticals: "If I do this now, what will result?" This leads directly to plans as a way of achieving goals.

6 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.

In the treatment of causality in Hobbs (2001), a distinction is made among the elements of a causal complex for an effect between those elements that *cause* the effect and those elements that merely *enable* the effect. Both imply a *causally-involved* relation.

Agents use their knowledge about causation and enablement to construct plans. In the classical 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.

$$
\begin{array}{l}
[goal'(g_2, a, e_2)\ \wedge\ know(a, cause(e_1, e_2))] \\
\quad \supset\ (\exists\, g_1)[goal'(g_1, a, e_1)\ \wedge\ cause(g_2, g_1) \\
\qquad\quad \wedge\, subgoal(a, e_1, e_2)]^3 \\
[goal'(g_2, a, e_2)\ \wedge\ know(a, enable(e_1, e_2))] \\
\quad \supset\ (\exists\, g_1)[goal'(g_1, a, e_1)\ \wedge\ cause(g_2, g_1) \\
\qquad\quad \wedge\, subgoal(a, e_1, e_2)]
\end{array}
$$

Backchaining on these axioms while instantiating them with axioms expressing causal knowledge results in hierarchical plans for achieving goals.

We can define a plan p by agent a to achieve goal g—$plan(p, a, g)$—as a composite entity with a set of subgoals and a set of subgoal relations among them. Then

[3] For notational convenience, we sometimes allow predications to be arguments of other predicates, where a proper treatment would reify the predications, as in Hobbs (1985). Also, defeasibility is not indicated in the axiom, but it could be, as in McCarthy (1980) or in Hobbs et al. (1993).

$\textit{subgoal-in}(e, p)$ says that eventuality e is one of the subgoals of plan p, and $\textit{subgoals-of}(p)$ is the set of subgoals of p.

It is sometimes said that it is a mystery where the goals come from. But it is easy to get around this difficulty by stipulating that agents have "thriving" as their top-level goal.

$$(\forall\, a)[\textit{agent}(a) \supset \textit{goal}(a, \textit{thrive}(a))]$$

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 and enabling altruistic behavior. Similarly, obliteration may be the best way of thriving in some circumstances. Thriving does not necessarily imply surviving.

Individual persons and individual computational agents are not the only kinds of planning 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 expected goals for a particular group of agents. 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 correlate with an agent's nationality, a role in an organization, a relationship, or a lifestyle choice, for example.

7 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 distinctions as subtle as those 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". This is a variety of envisionment. 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".

$$\begin{aligned}&\textit{decide-to}(a, e)\\ &\equiv \textit{change}(e_0, e_1) \wedge \textit{not}'(e_0, e_1) \wedge \textit{goal}'(e_1, e, a)\end{aligned}$$

That is, for a to decide to do e is for a to change from not having e as a goal to having e as a goal. The *decide-to* predicate is a bridge from the belief system to the planning module.

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.

We can posit a notion of "directly causes" (*dcause*) as a relation between an agent and an event, that is true when the agent causes the event with no intermediate, mediating events. Bodily events and some mental events are directly caused by the agent. So when a man moves his arm, he directly causes the arm to move. Knowledge about what kinds of agents can directly cause what kinds of events would be expressed in axioms of the form

$$p(a) \wedge q'(e, a, ...) \supset \textit{dcause}'(e_0, a, e)$$

That is, if an entity a is of type p (an agent) and e is a's doing something of type q, then there is the eventuality e_0 of a's directly causing e. For example,

$$\begin{aligned}&[\textit{person}(a) \wedge \textit{lift}'(e, a, x) \wedge \textit{arm}(x, a)]\\ &\supset (\exists\, e_0)\textit{dcause}'(e_0, a, e)\end{aligned}$$

Furthermore, if e actually occurs, then e_0 obtains as well; a directly caused p.

Direct causality is a variety of causality.

$$\textit{dcause}(a, e) \supset \textit{cause}(a, e)$$

In direct causality, there is no intermediate cause.

$$\begin{aligned}&\textit{dcause}(a, e)\\ &\supset \neg(\exists\, e_1)[\textit{cause}(a, e_1) \wedge \textit{cause}(e_1, e)]\end{aligned}$$

A plan is a way of manipulating the causal structure of the world to achieve one's goals. It must bottom out in events that are assured to occur at the proper time. There are three sources for this assurance:

1. The event is an action on the part of the agent that the agent is capable of carrying out at the required time, e.g., it can be directly caused when enabled.

2. The event will happen anyway at the required time, because of external causes.

3. The event is an action on the part of another agent, who is committed to perfoming the action at the required time.

For example, if you and I have a plan to start a car by together pushing it to the top of a hill and then letting it roll down and popping the clutch, I know the event of my pushing the car will happen because I can do that, I know the event of your pushing the car will happen because you have agreed to do it, and I know the car will roll down the hill because gravity will cause that to happen.

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.

$$\begin{aligned}&desire(a,e)\\&\equiv believe(a,part(e,e_3))\\&\quad\wedge believe(a,cause(e_3,thrive(a)))\end{aligned}$$

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. Similarly, desires don't *feel* like beliefs, but they can in part be modeled that way.

Planning can be viewed as 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 or are conditions that currently hold in the world.

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 made explicit a schedule for an agent's plan, we can talk about the agent's "schedule capacity" and "next unscheduled moment". A "deadline" is the time after which a plan no longer achieves it goal. 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.

8 Execution Envisionment and 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 can 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 developed 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 issues of aspect. We typically think of goals and subgoals as states. But corresponding to every such state, there are one or more actions or events that will bring it about. So we can refer to "executing a subgoal" where we mean carrying out an action that will lead to that subgoal being achieved. In general, an agent a executes an eventuality e in the service of a plan p at time t—$execute(a,e,p,t)$—whenever a is the agent of e, e is a subgoal in p, and e occurs at time t. An eventuality has been executed at time t if there is some previous time at which its execution took place.

$$\begin{aligned}&executed(e,p,t)\\&\equiv (\exists\, a,g,t_1)[plan(p,a,g)\wedge before(t_1,t)\\&\qquad\wedge execute(a,e,p,t_1)]\end{aligned}$$

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—$left\text{-}fringe(s,p)$.

The "remaining plan" p_1 of plan p at any given instant t is that part of the plan that has not yet been executed—$remaining\text{-}plan(p_1,p,t)$

When we make aspectual statements about events, we look at the fine structure of the event as it is being executed and say something about that execution. The fine structure of events is captured in plans, or in an equivalent causal structure. Therefore, this is the place in a theory of commonsense knowledge where a theory of aspect goes.

To start a plan is to execute an action in its left fringe.

$$\begin{aligned}&start'(e,a,p,t)\\&\equiv (\exists\, g,s,e_1)[plan(p,a,g)\wedge left\text{-}fringe(s,p)\\&\qquad\wedge member(e_1,s)\ \wedge\ execute'(e,a,e_1,p,t)]\end{aligned}$$

This says that e is a starting by agent a of plan p at time t if and only if p is a plan by a to achieve some goal g and e is the execution of an action e_1 by a at time t where e_1 is a member of the left fringe s of the plan.

To continue to execute a plan at time t is to execute an action the left fringe of the remaining plan at time t.

$$\begin{aligned}&continue'(e,a,p,t)\\&\equiv (\exists\, g,p_1,s,e_1)[plan(p,a,g)\\&\qquad\wedge remaining\text{-}plan(p_1,p,t)\\&\qquad\wedge left\text{-}fringe(s,p_1)\ \wedge\ member(e_1,s)\\&\qquad\wedge execute'(e,a,e_1,p,t)]\end{aligned}$$

To finish a plan for achieving a goal g is to execute the last action in the plan.

$$\begin{aligned}&finish'(e,a,p,t)\\&\equiv (\exists\, g,e_1,p_1)[plan(p,a,g)\\&\qquad\wedge remaining\text{-}plan(p_1,p,t)\\&\qquad\wedge subgoals\text{-}of(p_1)=\{e_1\}\end{aligned}$$

$$\wedge\, \mathit{execute}'(e,a,e_1,p,t)]$$

Stopping the execution of a plan at time t happens when an element of the plan was being executed before t and not after.

$$\begin{aligned}
&\mathit{stop}'(e,a,p,t)\\
&\equiv (\exists\, T_1,T_2)[\mathit{intMeets}(T_1,T_2)\\
&\qquad \wedge\, \mathit{endOf}(T_1)=t\\
&\qquad \wedge\, (\forall\, t\in T_1)(\exists\, e_1)[\mathit{subgoal\text{-}of}(e_1,p)\\
&\qquad\qquad \wedge\, \mathit{execute}(a,e_1,t)]\\
&\qquad \wedge\, (\forall\, t\in T_2)[\neg(\exists\, e_1)[\mathit{subgoal\text{-}of}(e_1,p)\\
&\qquad\qquad \wedge\, \mathit{execute}(a,e_1,t)]]]
\end{aligned}$$

It is possible for a plan to be stopped in the middle of its execution.

Once a plan is stopped, that precise plan may no longer be executable, because it may have specified scheduling decisions that can no longer be met. But it may be possible to achieve the same goal using different temporal parameters. In our full treatment, we define the relation *same-abstract-plan* between two plans if they differ only in their temporal parameters. (Ideally, we should extend this to plans that also differ in irrelevant choice of instruments.)

To resume a plan is to start a same abstract version of the remaining plan after some interval of time.

$$\begin{aligned}
&\mathit{resume}'(e,a,p,t)\\
&\equiv (\exists\, g,p_1,p_2,t_1)[\mathit{plan}(p,a,g)\ \wedge\ \mathit{stop}(a,p,t_1)\\
&\qquad \wedge\, \mathit{before}(t_1,t)\ \wedge\ \mathit{start}'(e,a,p_1,t)\\
&\qquad \wedge\, \mathit{same\text{-}abstract\text{-}plan}(p_1,p_2)\\
&\qquad \wedge\, \mathit{remaining\text{-}plan}(p_2,p,t_1)]
\end{aligned}$$

To restart a plan is to start it from the beginning after having stopped.

$$\begin{aligned}
&\mathit{restart}'(e,a,p,t)\\
&\equiv (\exists\, g,p_1,t_1)[\mathit{plan}(p,a,g)\ \wedge\ \mathit{stop}(a,p,t_1)\\
&\qquad \wedge\, \mathit{before}(t_1,t)\ \wedge\ \mathit{start}'(e,a,p_1,t)\\
&\qquad \wedge\, \mathit{same\text{-}abstract\text{-}plan}(p_1,p)]
\end{aligned}$$

To pause in a plan is to stop the execution of the plan, where there will be a resumption.

$$\begin{aligned}
&\mathit{pause}'(e,a,p,t)\\
&\equiv (\exists\, g,e_1)[\mathit{plan}(p,a,g)\ \wedge\ \mathit{stop}'(e,a,p,t)\\
&\qquad \wedge\, \mathit{resume}(a,p,t_1)\wedge \mathit{before}(t,t_1)]
\end{aligned}$$

A plan is ongoing for a time interval T if it is either being executed or there is a pause in its execution. That is, ongoing spans pauses.

$$\begin{aligned}
&\mathit{ongoing}(p,a,T)\\
&\equiv (\forall\, t)[\mathit{inside}(t,T)\\
&\qquad \supset\ [\mathit{continue}(a,p,t)\ \vee\ \mathit{pause}(a,p,t)]]
\end{aligned}$$

For an agent to suspend the execution of a plan is for the agent to stop executing the plan, where the agent has the goal of resuming it at some future time.

$$\begin{aligned}
&\mathit{suspend}'(e,a,p,t)\\
&\equiv (\exists\, g,e_1,t_1)[\mathit{plan}(p,a,g)\ \wedge\ \mathit{stop}'(e,a,p,t)\\
&\qquad \wedge\, \mathit{goal}(\mathit{resume}(a,p,t_1),a)\wedge \mathit{before}(t,t_1)]
\end{aligned}$$

The distinction between pausing and suspending is that with suspending there is only the intention to resume. The plan may not actually be resumed. Something is not a pause unless the resumption actually happens.

For an entity or event x to interrupt the execution of a plan is for x to cause the execution to be suspended.

$$\begin{aligned}
&\mathit{interrupt}'(e,x,a,p,t)\\
&\equiv \mathit{cause}'(e,x,\mathit{suspend}(a,p,t))
\end{aligned}$$

For an entity or event x to postpone the execution of a plan of agent a's from time t_1 to time t_2 is for x to cause a suspension in the plan where a has the goal of resuming at time t_2.

$$\begin{aligned}
&\mathit{postpone}'(e,x,a,p,t_1,t_2)\\
&\equiv [\mathit{cause}'(e,x,\mathit{stop}(a,p,t_1))\\
&\qquad \wedge\, \mathit{goal}(\mathit{resume}(a,p,t_2),a)\wedge \mathit{before}(t_1,t_2)]
\end{aligned}$$

To complete the execution of a plan is to execute every action in it.

$$\begin{aligned}
&\mathit{complete}(a,p,t)\\
&\equiv (\exists\, g)[\mathit{plan}(p,a,g)\\
&\qquad \wedge\, (\forall\, e)[\mathit{subgoal\text{-}in}(e,p)\\
&\qquad\qquad \supset\ (\exists\, t_1)[\mathit{before}(t_1,t)\ \vee\ t_1=t]\\
&\qquad\qquad\qquad \wedge\, \mathit{execute}(e,a,p,t_1)]]
\end{aligned}$$

The completion of a plan and its success are independent notions. We can execute all the actions in a plan and yet have the plan fail for reasons beyond our efforts. We can be in the midst of executing a plan and have the goal achieved, before we have actually performed all the actions in the plan.

To abort the execution of a plan is to stop the plan without completing it and with the intention not to restart or resume it.

$$\begin{aligned}
&\mathit{abort}'(e,a,p,t)\\
&\equiv (\exists\, g,e_1)[\mathit{plan}(p,a,g)\ \wedge\ \mathit{stop}'(e,a,p,t)\\
&\qquad \wedge\, \neg \mathit{complete}(a,p,t)\\
&\qquad \wedge\, (\forall\, t_1)[\mathit{before}(t,t_1)\\
&\qquad\qquad \supset\ \mathit{goal}(\mathit{not}(\mathit{resume}(a,p,t_1)),a)]]
\end{aligned}$$

9 Summary and Future Work

Part of the strong AI view of people is that they are planning mechanisms, and this is reflected as well in the way we talk about people's actions and in the strategies we devise for influencing the behavior of ourselves and other people. In the research we have described we have developed a more detailed formal realization of this picture than has heretofore been available. We have explicated at least one commonsense set of ideas about the structure of the mind, about how one thinks about past, present, future, and hypothetical situations, how this thinking is sometimes translated into action, and how actions are monitored and controlled.

Having completed the first sixteen of the thirty component theories of commonsense psychology, our first priority for future work is to complete the remaining fourteen theories. As with the first sixteen, we anticipate that many of the remaining theories will challenge some of the simplifying assumptions that have traditionally been made in formal knowledge representation research, and perhaps simplifying assumptions we have made as well. 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 in practical situations. We are particularly interested in validating these theories by using them to derive formal proofs of human strategies (Swanson and Gordon, 2005). 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 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 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.

Acknowledgements

The research described here was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). The views and conclusions contained herein are those of the authors and do not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred.

References

[1] Chisholm, Roderick M., 1957. *Perceiving: A Philosophical Study*, Cornell University Press, Ithaca, New York.

[2] Fikes, Richard, and Nils J. Nilsson, 1971. “STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving”, *Artificial Intelligence*, Vol. 2, pp. 189-208.

[3] Friedman, Nir, and Joseph Y. Halpern, 1999. ”Plausibility Measures and Default Reasoning: An Overview”, *Proceedings*, 14th Symposium on Logic in Computer Science, Trento, Italy, July 1999.

[4] Gettier, Edmund L., 1963. “Is Justified True Belief Knowledge?” *Analysis*, Vol. 23, pp. 121-123.

[5] Gordon, Andrew S., 2004. *Strategy Representation: An Analysis of Planning Knowledge*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

[6] Gordon, Andrew, and Jerry R. Hobbs, 2003. “Coverage and Competency in Formal Theories: A Commonsense Theory of Memory”, *Proceedings*, AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford, California, March 2003.

[7] Gordon, Andrew, and Jerry R. Hobbs, 2004. “Formalizations of Commonsense Psychology”, *AI Magazine*, Vol. 25, pp. 49-62.

[8] Gordon, Andrew, Abe Kazemzadeh, Anish Nair, and Milena Petrova, 2003. “Recognizing Expressions of Commonsense Psychology in English Text”, *Proceedings*, 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan, July 2003.

[9] Hobbs, Jerry R. 1985. “Ontological Promiscuity.” *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 61-69. Chicago, Illinois, July 1985.

[10] Hobbs, Jerry R., 2001. “Causality”, *Proceedings*, Common Sense 2001, Fifth Symposium on Logical Formalizations of Commonsense Reasoning, pp. 145-155, New York University, New York, New York, May 2001.

[11] Hobbs, Jerry R., in press. “Toward a Useful Concept of Causality for Lexical Semantics”, to appear in *Journal of Semantics.*

[12] Hobbs, Jerry R. and Feng Pan, 2004. “An Ontology of Time for the Semantic Web”, *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 1, March 2004, pp. 66-85.

[13] Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. “Interpretation as Abduction”, *Artificial Intelligence*, Vol. 63, Nos. 1-2, pp. 69-142.

[14] McCarthy, John, 1980. “Circumscription: A Form of Nonmonotonic Reasoning”, *Artificial Intelligence*, Vol. 13, pp. 27-39. Reprinted in M. Ginsberg, ed., *Readings in Nonmonotonic Reasoning*, pp. 145-152, Morgan Kaufmann Publishers, Inc., Los Altos, California.

[15] Moore, Robert C., 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.

[16] Narayanan, Srini, 1999. “Reasoning About Actions in Narrative Undertanding”, *Proceedings*, International Joint Conference on Artificial Intelligence, pp. 350-358, Stockholm, Sweden. Morgan Kaufmann, San Francisco, California.

[17] Schank, Roger, and Robert Abelson, 1977, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates. Inc., Hillsdale, New Jersey.

[18] Swanson, Reid, and Andrew Gordon, 2005. “Automated Commonsense Reasoning about Human Memory”, *Proceedings*, AAAI Spring Symposium on Metacognitive Computing, Stanford, California, March 2005.