The Theory of Mind in Strategy Representations

Andrew S. Gordon (gordon@ict.usc.edu)
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey CA USA

Abstract

Many scientific fields continue to explore cognition related to Theory of Mind abilities, where people reason about the mental states of themselves and others. Experimental and theoretical approaches to this problem have largely avoided issues concerning the contents of representations employed in this class of reasoning. In this paper, we describe a new approach to the investigation of representations related to Theory of Mind abilities that is based on the analysis of commonsense strategies. We argue that because the mental representations of strategies must include concepts of mental states and processes, the large-scale analysis of strategies can be informative of the representational scope of Theory of Mind abilities. The results of an analysis of this sort are presented as a description of thirty representational areas that organize the breadth of Theory of Mind concepts. Implications for Theory Theories and Simulation Theories of Theory of Mind reasoning are discussed.

Investigating the Theory of Mind

One of the most challenging areas of research in the cognitive sciences has concerned the Theory of Mind, in reference to the abilities humans have to perceive and reason about their own mental states and the mental states of other people. Along with the inherent difficulties in investigating behavior that is largely unobservable, researchers in this area are required to be extremely interdisciplinary. Many research fields contribute evidence that influences our understanding of these human abilities, although the methods used to gather this evidence are diverse.

Researchers in developmental psychology largely choose to investigate the Theory of Mind as a set of abilities that progressively emerge in normal child development (Wellman & Lagattuta, 2000). By the last half of their second year, toddlers demonstrate an understanding of the role of intentionality in action, and that other people have subjective experiences. By the age of four and five, children comprehend and use vocabulary to refer to mental states such as thoughts, imaginations, and knowledge. As children advance into grade-school years and adulthood, there is a growing appreciation of people as active constructors and interpreters of knowledge, and awareness that others have ongoing thoughts. There is evidence that Theory of Mind capabilities continue to improve into the later adult years, even while non-social reasoning abilities begin to degrade (Happé et al., 1998).

In the research area of abnormal psychology, compelling cases have been made relating illnesses such as autism (Baron-Cohen, 2000) and schizophrenia (Corcoran, 2001) to deficits in Theory of Mind abilities. Neuropathology studies of stroke patients have provided evidence that Theory of Mind mechanisms may be localized in the brain (Happé et al., 1999), and ongoing functional neuroimaging studies continue to provide further evidence for localization (Frith & Frith, 2000).

In search of a more process-oriented understanding of Theory of Mind abilities, it is the philosophy community that has made the most contributions, proposing two classes of process theories that have been extensively debated. First, the Theory Theory hypothesizes that Theory of Mind abilities are computed by prediction and explanation mechanisms by employing representation-level knowledge about mental attitudes (Gopnik & Meltzoff, 1997; Nichols & Stich, forthcoming). The opposing view is that of Simulation Theory (Goldman, 2000), which argues that Theory of Mind abilities are computed by imagining that you are in the place of the other person, then inferring their mental states by monitoring the processing that is done by your own cognitive mechanisms. While some high-level process-oriented cognitive models have been proposed (e.g. Nichols and Stich, 2000), there are many unanswered questions that prohibit the creation of detailed, computational models of Theory of Mind abilities.

Most lacking in our theoretical understanding of Theory of Mind abilities is a description of the specific *contents* of the mental representations that are employed in this reasoning. There is general agreement that these representational elements must include concepts such as *beliefs* and *desires* (e.g. Harris, 1996), and these two concepts in particular have taken a privileged role in the cognitive models that have been proposed. A potential benefit of the focus on these concepts is that this representational area (beliefs, desires, intentionality) is among the very few where established axiomatic theories have been developed in the artificial intelligence community (Cohen & Levesque, 1990). Continued artificial intelligence progress in developing axioms for inference concerning mental states (e.g. Ortiz, 1998) will greatly support the plausibility of the Theory Theory approach.

However, there is a general sense throughout the fields investigating Theory of Mind abilities that the contents of these representations go far beyond simple notions of beliefs and desires, particularly among developmental psychologists investigating the role that language plays in acquiring mental state concepts. Several studies have been conducted that investigate the linguistic environment of children for the presence of Theory of Mind related terms, where the conceptual scope is much more broadly construed. Dyer et al. (2000) best exemplifies the broad conceptual scope of this line of work, which compared the frequency of 455 mental state terms that appear in young children's storybooks. This list included 102 cognitive state terms (e.g. notice, wonder), 152 emotional state terms (e.g. nervous, boring), 84 desire and volition terms (e.g. hope, wish), and 117 moral evaluation and obligation terms (e.g. ought, terrible), where the complete list was compiled from previous language studies.

While these linguistic approaches help to broaden our conception of the scope of representational elements in Theory of Mind reasoning, many of the traditional concerns about the relationship between language and mental representation may apply. Particularly, there is no reason to believe that any full enumeration of mental state terms must parallel the breadth of concepts that are represented and manipulated by reasoning processes. The inherent subjectivity of these concepts may serve to restrict the introduction of new vocabulary in the lexicon as compared with other topics of discourse. Likewise, the remarkable creativity that is evident in human language use may mislead us to believe that there are representational distinctions between concepts that are in fact functionally synonymous. While these linguistic approaches have been persuasive in arguing for a broader scope of Theory of Mind representations in our cognitive models, a new investigative methodology for concept enumeration would be useful.

Analogy as an Investigative Tool

In previous work (Gordon, 2001a), we argued that progress in a different area of cognition – that of analogical reasoning – could be a basis for a novel methodology for the investigation of mental representations. As a cognitive process, analogical reasoning has received an enormous amount of attention, both theoretical and experimental, with the aim of understanding how people draw analogies between two different cases in working memory. The prevailing explanation is based on the notion of structural alignment of the mental representations that people have of these cases (Gentner, 1983). That is, two different cases are judged as strongly analogous when portions of the structured mental representation of one case can be mapped onto structurally identical portions of the other. Strong empirical support for the structure mapping theory of analogical reasoning (see Gentner & Markman, 1997, for a review) presents an opportunity: if structural alignment of representations are necessary to process analogies, then an analysis of the analogies that people naturally make can reveal the sorts of representations that they must employ.

In this previous work, two main claims were put forth. First it was noted that there is something particularly interesting about the commonsense notion of a strategy as it relates to analogies between planning cases. People readily see analogies between planning behaviors exhibited in vastly different goal-driven domains. For example, a retreating military force that destroys the supplies that they can't take with them may be viewed as analogous to the company that publicly releases its closely guarded industrial secrets in the face of a hostile corporate takeover. In both cases we would say the actors were using the same strategy, one that is so commonly recognized that it has been given a name, *scorched earth policy*. In accordance with the structure mapping theory of analogy, it was argued that strategies like this one are structured mental representations that are shared between the analogous planning cases where they are employed.

The second claim was that mental representations of strategies necessarily include references to the mental states and processes of people. For example, to be considered as an example of scorched earth policy, it must be the case that the actor foresees he will loose possession of a valuable resource to an advancing enemy, he foresees that after the enemy gains possession of it he will use the resource to further advance against him, and that the actor imagines that what he does to these resources will make them useless to the enemy. Strategic analogies show us that concepts such as these that specifically refer to mental processes must be explicitly represented in cognition. As the mental state concepts in these statements are exactly the sort relevant to Theory of Mind abilities, our claim is that the analysis of strategies provides a means of identifying the breadth of reified mental state concepts that are available in support of this class of reasoning.

In order to explore the representational scope of strategies, we undertook a large-scale strategy representation effort (Gordon, 2001b). First, 372 commonsense strategies were collected from 10 different planning domains using directed expert interviews, the analysis of texts that are encyclopedic of strategies in a particular domain, and the introspective elaboration of strategies in our own areas of expertise. To identify the representational requirements of this catalog of strategies, we developed a notational form called a *pre-formal representation* that would allow us to commit to the specific semantic elements in the representation of a strategy without adhering to the syntactic constraints that would be necessary in more formal, logic-based representations. After authoring pre-formal representations of each of the 372 strategies, the component concepts were grouped into sets of synonyms to form a controlled vocabulary consisting of 989 unique concepts. This list was then organized into 48 representational areas that parallel both those that are traditionally the subject of formal commonsense knowledge representation (e.g. time and events) and those that are viewed as component cognitive processes in previous cognitive modeling work (planning and memory retrieval).

Eighteen of the representational areas that were identified in this previous work did not concern the mental states and processes of people. A large portion of these areas related more generally to the physical world, including concepts of time, space, events, states, objects, numbers, sets, and taxonomies. The remaining portion of these eighteen areas concerned people directly, but not their mental states in particular, and included terms for the relationships they hold, the organizations they participate in, their abilities, activities, and non-mental actions.

The other thirty representational areas that were identified deal specifically with the mental life of people. What is interesting about this collection of representational terms is that its scope is significantly larger than what has been suggested in cognitive models of Theory of Mind abilities or even in the contents of the lexicons used in the analysis of language for Theory of Mind concepts.

The primary direction in which these representational areas expand the scope of previous work is with respect to folk psychological conceptions of mental *processes*, whereas previous work has focused mostly on mental *states*. While the terms revealed in our investigation certainly include mental state concepts such as beliefs and desires, these are coupled with concepts describing the mental processes that affect these states, such as the mental processes of removing the justification for a belief and the process of abandoning of a goal to achieve some desired state. In short, the representations that appear to be necessary to account for strategic analogies outline a set of processes that constitute a cognitive architecture.

Theory of Mind Representations

In order to elaborate on the mental state and mental process components that are evident in the organization of strategy representation terms, this section briefly describes each of the thirty representational areas (of the 48 total) specifically related to Theory of Mind reasoning. Each area is listed with a short area title, the number of unique representational terms (out of 989) in the area found in strategy representations, a short definition of the scope of the area, and a few examples of the specific terms in the area.

1. Managing knowledge (30 terms): The knowledge that agents have is a set of beliefs that may be true or false based on certain justifications, and can be actively assumed true, affirmed, or disregarded entirely. Examples: *Assumption*, *Justification*, *Revealed false belief*.

2. Similarity comparison (16 terms): Agents can reason about the similarity of different things using different similarity metrics, where analogies are similar only at an abstract level. Examples: *Class similarity*, *Similarity metric*, *Make analogy*.

3. Memory retrieval (3 terms): Agents have a memory that they use to store information through a process of memorization, and may use memory aids and cues to facilitate retrieval. Examples: *Memory cue*, *Memory retrieval*, *Memorize*.

4. Emotions (8 terms): Agents may experience a wide range of emotional responses based on their appraisal of situations, which defines their emotional state. Examples: *Anxiety emotion*, *Pride emotion*, *Emotional state*.

5. Explanations (17 terms): Agents generate candidate explanations for causes in the world that are unknown, and may have preferences for certain classes of explanations. Examples: *Candidate explanation*, *Explanation preference*, *Explanation failure*.

6. World envisionment (48 terms): Agents have the capacity to imagine states other than the current state, to predict what will happen next or what has happened in the past, and to determine the feasibility of certain state transitions. Examples: *Causal chain*, *Envisioned likelihoood*, *Possible envisioned state*.

7. Execution envisionment (23 terms): One mode of envisionment is that of imagining the execution of a plan for the purpose of predicting possible conflicts, execution failures, side effects, and the likelihood of successful execution. Examples: *Envisioned failure*, *Side effect*, *Imagine possible execution*.

8. Causes of failure (31 terms): In attempting to explain failures of plans and reasoning, agents may employ a number of explanation patterns, such as explaining a scheduling failure by the lack of time, or a planning failure by a lack of resources. Examples: *False triggered monitor*, *Lack of ability*, *Successful execution of opposing competitive plan*.

9. Managing expectations (8 terms): Envisionments about what will happen next constitute expectations, which can be validated or violated based on what actually occurs. Examples: *Expectation violation*, *Unexpected event*, *Remove expectation*.

10. Other agent reasoning (8 terms): Envisionments about the planning and reasoning processes of other

agents allow an agent to imagine what they would be thinking about if they were them. Examples: *Guess expectation*, *Guess goal*, *Deduce other agent plan*.

11. Threat detection (15 terms): By monitoring their own envisionments for states that violate goals, an agent can detect threats and track their realization. Examples: *Envisioned threat*, *Realized threat*, *Threat condition*.

12. Goals (27 terms): Goals of agents describe world states and events that are desired, and include both states and events that are external to the planner as well as those that characterize desired internal mental states and processes. Examples: *Auxiliary goal*, *Knowledge goal*, *Shared goal*.

13. Goal themes (6 terms): A potential reason that an agent may have a goal could be based on the roles that agents have in relationships and organizations, or because of a value that they hold. Examples: *Generous theme*, *Good person theme*, *Retaliation theme*.

14. Goal management (28 terms): Agents actively manage the goals that they have, deciding when to add new goals, commence or suspend the pursuit of goals, modify or specify their goals in some way, or abandon them altogether. Examples: *Currently pursued goal*, *Goal prioritization*, *Suspend goal*.

15. Plans (32 terms): The plans of agents are descriptions of behaviors that are imagined to achieve goals, and can be distinguished by the types of goals that they achieve or by how they are executed, and may be composed of other plans or only partially specified. Examples: *Adversarial plan*, *Repetitive plan, Shared plan*.

16. Plan elements (28 terms): Plans are composed of subplans, including branches that are contingent on factors only known at the time of execution. They may have iterative or repetitive components, or include components that are absolutely required for a plan to succeed. Examples: *If then*, *Iteration termination condition*, *Triggered start time*.

17. Planning modalities (17 terms): The selection of plans can be done in a variety of different ways, such as adapting old plans to current situations, collaboratively planning with other agents, and counterplanning against the envisioned plans of adversaries. Examples: *Adversarial planning*, *Auxiliary goal pursuit*, *Imagined world planning*.

18. Planning goals (27 terms): The planning process is directed by abstract planning goals of an agent, which include goals of blocking threats, delaying events, enabling an action, preserving a precondition, or satisfying the goals of others. Examples: *Avoid action*, *Delay duration end*, *Maximize value*.

19. Plan construction (30 terms): Agents construct new plans by specializing partial plans, adding and ordering subplans, and resolving planning problems when they arise. Examples: *Candidate plan*, *Planning failure*, *Planning preference*.

20. Plan adaptation (18 terms): Existing plans can be adapted and modified by substituting values or agency, and by adding or removing subplans to achieve goals given the current situation. Examples: *Adaptation cost*, *Adaptation failure*, *Substitution adaptation*.

21. Design (8 terms): One modality of planning is design, where the constructed plan is a description of a thing in the world within certain design constraints, and where the resulting things have a degree of adherence to this design. Examples: *Design adherence*, *Design failure*, *Designed use*.

22. Decisions (38 terms): Agents are faced with choices that may have an effect on their goals, and must decide among options based on some selection criteria or by evaluating the envisioned consequences. Examples: *Best candidate*, *Decision justification*, *Preference*.

23. Scheduling (23 terms): As agents select plans, they must be scheduled so that they are performed before deadlines and abide by other scheduling constraints. Plans may have scheduled start times and durations, or may be pending as the planner waits for the next opportunity for execution. Examples: *Deadline*, *Pending plan*, *Scheduling constraint*.

24. Monitoring (18 terms): Agents monitor both states and events in the world and in their own reasoning processes for certain trigger conditions which may prompt the execution of a triggered action. Examples: *First monitor triggering*, *Monitoring duration*, *Monitor envisionment*.

25. Execution modalities (11 terms): Plans can be executed in a variety of ways, including consecutively along with other plans, in a repetitive manner, and collaboratively along with other agents. Examples: *Concurrent execution*, *Continuous execution*, *Periodic execution*.

26. Execution control (28 terms): A planner actively decides to begin the execution of a plan, and may then decide to suspend or terminate its execution. A suspended plan can later be resumed from the point the agent left off. Examples: *Execution delay*, *Suspend execution*, *Terminate activity*.

27. Repetitive execution (16 terms): Some plans and subplans are executed iteratively for some number of times, or repetitively until some termination condition is achieved. Examples: *Current iteration*, *Iteration completion*, *Remaining repetition*.

28. Plan following (29 terms): Agents track the progress of their plans as they execute them in order to recognize when deadlines are missed, preconditions are satisfied, and when they have successfully achieved the goal. Examples: *Achieve precondition*, *Miss deadline*, *Successful execution*.

29. Observation of execution (29 terms): Agents can track the execution of plans by other agents, evaluating the degree to which these executions adhere to performance descriptions known to the observing agent. Examples: *Observed execution*, *Assessment criteria*, *Performance encoding*.

30. Body interaction (15 terms): The physical body of an agent translates intended actions into physical movements, and sometimes behaves in unintended ways. The body modifies the planner's knowledge through perception of the world around it, and by causing a sensation of execution. Examples: *Impaired agency*, *Nonconscious execution*, *Attend*.

Discussion

There exists no infallible technique for identifying the contents of the mental representations used in reasoning. The approach described here, where our theoretical understanding of analogical reasoning is used as an investigative tool, relies heavily on our analytic abilities in describing the shared relational structure of analogous cases as much as on the validity of the structure-mapping theory itself. We feel that while the specific concepts chosen in the course of authoring pre-formal representations of 372 strategies can rightly be questioned, the scope of these concepts as a whole cannot. The evidence provided by terms used in strategy representations suggests that the scope of mental representations that may support Theory of Mind abilities includes concepts for both the mental states of people and of the cognitive processes that they employ.

This evidence of process-oriented mental representations does not by itself provide support for either of the two prominent Theory of Mind theories (Theory Theory and Simulation Theory). However, it does have relevance to how proponents of these theories proceed to produce more detailed, even computational, process models of Theory of Mind abilities.

Proponents of the Theory Theory should view these representation areas as a catalog of the component theories that will be necessary to specify a complete folk psychology, in much the same way that artificial intelligence researchers have attempted to define the component theories of naïve physics (Hayes, 1985). If the two endeavors are indeed similar, then terms like those that comprise the representational areas listed here will appear as notations (predicates or otherwise) in formal axiomatic theories that could drive deductive reasoning. Because breadth of component theories for a full folk psychology appears to be at least as rich as those in naïve physics, we would expect that the same methodological problems in specifying these theories would prohibit progress (see Davis, 1998). As folk psychology has received little attention within the artificial intelligence community as compared with naïve physics, few axiomatic theories exist today for the majority of the representational areas that are listed in this paper.

While axiomatic theories have not been forthcoming, most of these representational areas have been extensively studied as cognitive processes. Cognitive science and artificial intelligence researchers have constructed an enormous number of computational models in support of our theoretical understandings, with one notable exception of Theory of Mind reasoning itself (representational area 10, Other Agent Reasoning). For proponents of the Simulation Theory it is this set of computational models that will have to be employed in the off-line reasoning that allows a person to perform Theory of Mind tasks. Evidence of mental representations that correspond to these processes could suggest that there is a representational interface to support off-line reasoning. That is, terms like those in each of the representational areas could be viewed as a vocabulary for expressing inputs (e.g. commands and arguments) to these processes as well as their outputs (e.g. inferences). Further agreement within the cognitive modeling community concerning inputs and outputs could potentially promote the development of more modular computational theories, facilitating the integration of models that will be necessary in providing a process account of Simulation Theory, among others.

While the investigation of Theory of Mind representations may affect the theoretical debate only in the long run, its utility in linguistic studies of Theory of Mind language use may be more direct. Specifically, the identification of these representation areas – as well as the specific terms that appear in each – may be valuable in identifying a broader lexicon for use in the analysis of language data. For example, a process-oriented term such as *suspend goal* (from area 14, Goal Management) is expressible in a wide variety of ways in English, as in "Let's *put it off for now*" or "I'll *come back to it later*" where the direct object in both statements is the suspended goal. Compiling word and phrase lexicons for each of the terms in these representational areas could provide enough coverage over a language to facilitate more automated text analysis approaches, which in turn could greatly scale up the amount of linguistic data that could be analyzed.

Conclusions

While there has been great interest in understanding the Theory of Mind abilities of people, the experimental and theoretical approaches to this problem have largely avoided issues concerning the *contents* of representations employed in this class of reasoning. In this paper, we have argued that progress in our understanding of a different cognitive process – that of analogical reasoning – provides us with a tool that can be used to investigate these representations in a new way. The curious nature of commonsense strategies, in accounting for analogies in planning domains and including references to mental states and processes, makes them a particularly important subject of analysis. By conducting a large-scale analysis of strategies from many planning domains, authoring pre-formal representations for each, we have improved the understanding of the scope of representations that would be available in support of Theory of Mind reasoning abilities. In addition to the mental state concepts that have traditionally been discussed in Theory of Mind research, this investigation suggests that rich representations of mental processes are also part of our representations.

References

Baron-Cohen, S. (2000) Theory of mind and autism: a fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Cohen, P. and Levesque, H. (1990) Intention is Choice with Commitment. *Artificial Intelligence* 42, 213-261.

Corcoran, R (2001) Theory of Mind in Schizophrenia. In: D. Penn and P. Corrigan (Eds.) *Social Cognition in Schizophrenia*. APA.

Davis, E. (1998) The Naïve Physics Perplex, *AI Magazine*, Winter 1998.

Dyer, J., Shatz, M., & Wellman, H. (2000) Young children's storybooks as a source of mental state information. *Cognitive Development* 15, 17-37.

Frith, C. & Frith, U. (2000) The physiological basis of theory of mind: functional neuroimaging studies. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Gentner, D. (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7, 155-170.

Gentner, D. & Markman, A. (1997) Structure mapping in analogy and similarity. *American Psychologist* 52, 45-56.

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

Gopnik, A. & Meltzoff, A. (1997). Words, thoughts, and theories. Cambridge, Mass.: Bradford, MIT Press.

Gordon, A. (2001a) Strategies in Analogous Planning Cases. In J. Moore & K. Stenning (eds.) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gordon, A. (2001b) The Representational Requirements of Strategic Planning. Fifth symposium on Logical Formalizations of Commonsense Reasoning. (http://www.cs.nyu.edu/faculty/davise/commonsense01/).

Happé, F., Brownell, H., & Winner, E. (1998) The getting of wisdom: Theory of mind in old age. *Developmental Psychology*, 34 (2), 358-362.

Happé, F., Brownell, H., & Winner, E. (1999) Acquired 'theory of mind' impairments following stroke. *Cognition* 70, 211-240.

Harris, P. (1996) Desires, beliefs, and language. In P. Carruthers and P. Smith (Eds.) *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Hayes, P. (1985) The second naïve physics manifesto. In J. Hobbs & B. Moore, Formal Theories of the Commonsense World. Ablex Publishing.

Nichols, S. & Stich, S. (2000) A cognitive theory of pretense. *Cognition* 74, 115-147.

Nichols, S. & Stich, S. (forthcoming) How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In Q. Smith and A. Jokic (Eds.) *Consciousness: New Philosophical Essays*, Oxford University Press.

Ortiz, C. (1999) Introspective and elaborative processes in rational agents. *Annals of Mathematics and Artificial Intelligence* 25, 1-34.

Wellman, H.M., & Lagattuta, K. H. (2000). Developing understandings of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.