Psychological Simulation and Beyond *

Ian Pratt
Department of Computer Science,
University of Manchester,
Manchester, M13 9PL, UK
ipratt@uk.ac.man.cs.ux

Abstract

In this paper, we examine the suggestion that inferences about another person's state of mind can proceed by *simulation*. According to that suggestion, one performs such reasoning by *imagining* oneself in that person's state of mind, and *observing* the evolution of that imagined cognitive state. However, this simulation-based theory of psychological inference suffers from a number of limitations. In particular, whilst one can perhaps observe the probable *effects* of an given cognitive state by putting oneself in that state, one cannot thus observe its probable *causes*. The purpose of this paper is to propose a solution to this problem, within the spirit of the simulation-based theory of psychological inference. According to the *indexing thesis*, certain cognitive mechanisms required for non-psychological inference can be re-used for hypothesising psychological causes. The paper concludes by discussing some of the possible implications of the indexing thesis.

1 Psychological inference

Let us begin with an example. Suppose a company employee E is told by his colleague F to take some coffee to a certain office where a meeting has been taking place. Suppose further that E knows that the meeting has just broken up. Then if E is minimally intelligent and co-operative, he will not blindly obey the instruction, but will pass his information on to F. In doing so, E has inferred the following: (i) that F *wants* coffee to be taken to the office; (ii) that F wants this because F has the *goal* of giving the people in the meeting coffee; and (iii) that F *believes* (falsely) that the people at the meeting are still in the office. We call inferences involving the *goals*, *plans*, *beliefs*, *etc.* of other agents *psychological inferences*.

In this paper, we examine the suggestion that psychological inference can proceed by *simulation*. According to that suggestion, one reasons about the state of mind of another

*This work is supported in part by the Joint Councils' Research Initiative in Cognitive Science. MRC grant number SPG8920254.

person by *imagining* oneself in that person's situation, and *observing* the evolution of one's imagined cognitive state. However, this simulation-based theory of psychological inference suffers from a number of problems. The purpose of this paper is to examine one of those problems, and to suggest a solution within the spirit of the original theory.

2 Background

Within artificial intelligence, research into psychological inference has fallen into two main traditions. The first, arising from the work of the Yale school on story-understanding (e.g. Wilensky[14]) focuses primarily on *plan-attribution*. On this approach, inferring the plan behind an agent's actions involves the repeated application of special rules (sometimes called 'reverse planning rules') which allow the system to pass from those actions to the goals that they facilitate or achieve. Recent work in this tradition includes that of Allen[1] and Litman and Allen[10] on plan-attribution in human dialogue.

The second tradition of psychological inference within artificial intelligence uses *epistemic logics* to reason about belief and knowledge, whilst tending to neglect goals, plans, intentions, etc. (see Konolige[9] for a survey). These epistemic logics themselves fall into two distinct classes, corresponding to two competing philosophical accounts of belief and knowledge. *Propositional logics*, following the work of Hintikka[7], take the objects of belief to be propositions—equivalently, sets of possible worlds. In thus characterising belief as a relation between a believer and a set of possible worlds, it is natural for such logics to idealise the rationality of agents by taking belief to be both consistent and closed under logical consequence[1]. The second class of epistemic logics, the *sentential logics*, take the objects of beliefs to be sentence-like entities, sentences, as it were, in the agent's language of thought. Sentential logics appear to be more promising for modelling sub-rational agents, in particular, agents whose beliefs may be inconsistent or not logically closed. Included in the category of sententialist logics, and of particular relevance to this paper, is the work of Haas[6], who, following Creary[4], takes psychological inference to proceed by *simulation*.

3 The simulation thesis

The simulation-based theory of psychological inference adopts the thesis that psychological inference is sometimes a matter of simulation (hereinafter, the *simulation thesis*). The idea can be illustrated with an example taken from chess. In deciding on a move in a chess game, I must reason about which threats will be obvious and which difficult to see until it is too late. According to the simulation thesis, to decide whether a given threat will be obvious

[1] Some epistemic logicians have worked on overcoming these idealisations (see, for example Levesque[11] and Fagin and Halpern[5]). Other researchers, however—particularly within the philosophical community—regard consistency and closure as desirable properties for any reconstruction of the notions of belief and knowledge, not as mere idealisations whose sole purpose is to render epistemic logics simple and perspicuous. See, for example, Stalnaker[13].

to my opponent, I imagine myself in his position, and observe how easily I can, within that imagined cognitive predicament, discover the threat for myself. That is, I imagine myself knowing what my opponent knows (the positions of the pieces on the board, but not the other player's immediate plans), and desiring what my opponent desires (that he should win), and I *simulate* the processes that may be presumed to be occurring in his mind.

Notice that the fact that I actually know the purpose of my chess gambit is nothing to the point. In imagining myself in my opponent's position, I temporarily suspend this knowledge, working only from what I think my opponent knows. In doing so, I automatically simulate the time- and space-constraints under which my opponent is labouring, for the simple reason that, in performing the simulation, *I* am labouring under similar constraints. Thus, the simulation-based theory of psychological inference prides itself on the natural way in which it yields conclusions about agents whose reasoning powers are sub-optimal: predictions of sub-optimality due to limitations of time, space and knowledge arise naturally from the fact that whoever is doing the prediction will be labouring under such constraints himself.

I suggest that psychological reasoning by psychological simulation occurs often in everyday situations. Why might I worry that my friend may doubt that I like him? Because I have, unavoidably, missed several recent appointments with him, uttered comments which, on reflection, might have been misinterpreted, and so forth. I know that the comments might have been misinterpreted because I can imagine hearing them (and can imagine having the particular concerns and interests with which I know my friend to be currently preoccupied), whereupon the offense-giving interpretations, which had not occurred to me before, now become obvious. I know that these interpretations, together with the missed appointments and so forth, might cause my friend to doubt my loyalty because, as I imagine having experienced the things I believe him to have experienced, that thought occurs to me.

Again, the fact that I actually know that I still like my friend, or that I can devise a clever argument which would convince my friend that his doubts are unjustified, are nothing to the point. It is enough that, in contemplating the situation from my friend's point of view, I can experience coming to a particular conclusion. To repeat: the simulation-based theory of psychological inference prides itself on the natural way in which it yields conclusions about agents whose reasoning powers are sub-optimal.

In reasoning about another person's reasoning by imagining oneself in that person's position, one is exploiting the fact that, in one's own brain, one has an analogue model of another person's brain: the ability to imagine oneself in alternative cognitive predicaments is the means by which one can compensate for the differences between one's own thinking and that of the agent one is simulating. Notice that, in using psychological simulation to predict that another person might well fail to see a threat in chess, or fail to make an inference, or otherwise engage in some seductive but erroneous piece of thinking, one need not have any sort of *theory* of how that person thinks. Such knowledge is unnecessary because psychological inference can simply trade on the fact that one person's cognitive mechanisms are really very much like another's. For instance, when I predict my opponent's likely responses to a chess gambit, I use many of those cognitive mechanisms which enable me to reason non-psychologically about chess. Whatever cognitive mechanisms that allow me to scan a board for threats and opportunities, compute the values of exchanges of pieces and so forth, now find alternative employment: they enable me to make inferences about how my opponent will or may think in a given situation. I do not need extra axioms or rules which describe how I do these things, or which tell me how long I will take to do them: such axioms and rules would at best amount to a pointless duplication.

4 A problem with the simulation-based theory

The simulation-based theory of psychological inference has a long history, going back at least to Hobbes[8]. More recently, interest in simulation was re-kindled by Kenneth Craik[3], and similar ideas have been discussed in a variety of contexts in artificial intelligence. (For a survey, see Pratt[12].) Although simple in outline, however, the simulation thesis quickly leads to considerable difficulties which must be addressed if the simulation-based theory is to be developed.

Here is one such difficulty. The chess example illustrates how, if one imagines oneself in a given cognitive predicament, one can observe the way in which that predicament will evolve. But whilst one can thus observe what *later* states a given cognitive predicament would normally *give rise to* (i.e. what *effects* it would normally have), one cannot observe what *earlier* states would normally *have given rise to* that cognitive predicament (i.e. what *causes* it would normally have). Yet many of the psychological inferences we want to perform involve inferring not the *effects* of given psychological states, but their *causes.* So it is with the example of section 1: when E attributes to F the goal of giving the people in the meeting coffee, E is making an inference (to us, an obvious inference) about what earlier psychological state *caused* F's decision to ask E to take the coffee to the office. And it is difficult to see how such an inference could proceed by simulation: by definition, psychological simulation can only run forward in time.

To be sure, if one has a limited number of specific hypotheses about the psychological causes of an action, one may test those hypotheses by simulation: simply by putting oneself into each of the hypothesised psychological states in turn, and seeing if the results are consistent with one's current view of the agent one is simulating. But simulation can be of no help in *generating* a suitably constrained set of hypotheses in the first place. The inability of simulation to generate hypotheses about the causes of psychological states constitutes a major limitation of the simulation-based view of psychological inference. Apparently, something more than psychological simulation is required for effective psychological inference.

5 The indexing thesis

I come now to examine a partial solution to the problem of generating hypotheses about the causes of psychological states, a solution which I take to preserve the spirit of the simulation-based theory. The current status of the solution is purely that of a proposal. However, in the next section, I explore some of its implications with a view to its empirical testability.

The basic idea can be illustrated by comparing the following two inferential tasks.

> (1) Imagine yourself in a room with some very high shelves and a large wooden chest, and consider the action of dragging the chest over to the shelves. If you now ask what potentially useful goal this action might enable you to achieve, the answer will be obvious: that of reaching the shelves.
>
> (2) Imagine yourself in the same room, but now suppose that someone tells you to put the chest beneath the shelves. If you now ask yourself why this person intends that the chest should be placed there, the answer will again be obvious: presumably, that person wants to be able to reach the shelves.

Prima facie, inferences (1) and (2) are very different. In inference (1), one thinks up a useful goal given a condition (or given an action to achieve that condition) which makes it possible to realise that goal; in inference (2), by contrast, one hypothesizes the cause, in terms of a higher-level goal, of someone's intention to achieve a given state. Inference (1) is not at all psychological (there is no reasoning about anyone's reasoning); inference (2), by contrast, is a centrally about the psychological cause of a psychological state. Let us call the first mode of reasoning, in which the problem is simply to think of a goal G that a given condition P enables or facilitates, *goal-activation*; let us call the second mode of reasoning, in which the problem is to hypothesise a possible goal G as a cause of an agent S's intention to bring about P, *goal-attribution*. Goal-activation has a forward-chaining character. One starts with some state P and thinks up a goal that P will help achieve. Goal-attribution, by contrast, is often a matter of *reasoning about* S's backward-chaining inference. One supposes S to have started with some goal G and, as a result, to have adopted some plan which involves achieving P; the problem is: given P, solve for G.

But in spite of their differences, these two modes of reasoning have a similar character. The crux of the similarity is this: in both cases, the reasoner must rely on some cognitive mechanism whereby goals are *indexed* by conditions or actions that enable or facilitate them. For example, in both inferences (1) and (2) above, the reasoner must be able to think up the goal of reaching the shelves by considering the state of affairs in which the chest is beneath the shelves. Now, a plausible thesis is that, fundamentally, the same goal-indexing mechanisms are used in *goal-attribution* as are used in corresponding cases of *goal-activation*. In particular, the same indexing mechanisms are used in inferences (1) and (2) above. Let us call this thesis the *indexing thesis.*

When one is faced with the problem of looking for the cause of an agent's intention to achieve some state of affairs P, and when one is considering the possibility that that intention may have arisen through a process of backward-chaining planning from some goal G, then, in seeking possible candidates for G, one performs forward-chaining reasoning starting from P, looking for sensible goals that P might help achieve. Put loosely, one might say that, according to the indexing thesis, the cognitive mechanisms one uses in (non-psychological) forward-chaining inference get re-used in reasoning about the backward-chaining inference of other agents. Though the indexing thesis and simulation thesis are logically independent of each other, they are nevertheless in much the same spirit. For, according to both theses, cognitive mechanisms needed anyway for wholly non-psychological inference find alternative employment in psychological inference. However, whilst the simulation thesis is about how one infers the probable (psychological) *effects* of a given psychological state, the indexing thesis is about how one hypothesises some possible (psychological) *causes* of a given psychological state. Thus, the indexing thesis complements the simulation thesis.

6 Implications

We can get some idea of the implications of the indexing thesis by considering the effect on goal-activation and goal-attribution of one's knowledge of whether various conditions obtain that affect the achievability of certain goals. These implications suggest that the indexing thesis is empirically testable, as I shall explain.

It is important to realise that, in goal-activation, a new (or newly perceived) condition P may suggest a goal G even if G is not directly attainable as things stand. Consider, for example, a variant of inference (1), in which the shelves can only be reached by placing a stool on top of the chest. In that case, the action of pushing the chest to beneath the shelves may still suggest the goal of reaching the shelves, even though achieving that goal may require an extra action (putting the stool on top of the chest). Clearly, if goal-activation is to be useful, a situation P (or an action that achieves it) must sometimes suggest a goal G even if G requires the achieving of further preconditions which are not as yet satisfied. (Indeed, such preconditions may turn out to be unsatisfiable.) On the other hand, our knowledge of what conditions obtain must constrain the the process of goal-activation to *some* extent: otherwise there is no limit to the hopeless goals that any condition P might suggest. Ideally, the less likely G is to be achievable given P, the less likely a reasoner should be to think of G when he considers what goals P might facilitate. Thus, in goal-activation, the indexing of a goal G by a precondition P should be sensitive to one's beliefs concerning the various other preconditions of G[2].

Now, the indexing thesis states that the same goal-indexing mechanisms underly both goal-activation and goal-attribution. But we just argued that, in goal-activation, the indexing of a goal G by a precondition P should be sensitive to one's beliefs concerning the various other preconditions of G. If so, then, we should expect that in goal-*attribution*, the indexing of a goal G by a precondition P will be similarly sensitive to one's beliefs concerning the various other preconditions of G.

The question arises as to whether this consequence of the indexing thesis desirable. Now, in many goal-attribution tasks (e.g., the example of section 1), one must be prepared to attribute to an agent a goal G which only makes sense on the supposition that that agent has a false belief about some precondition for achieving G. Indeed, such cases of goal-attribution are often essential for detecting these false beliefs. But then, one needs a mechanism for thinking up a goal G given a precondition P, even if, according to *one's own* knowledge,

[2] For brevity's sake I speak of the *preconditions of goal G* rather than the *preconditions of actions which will achieve goal G*.

goal G has some unsatisfied (or even unsatisfiable) preconditions. On the other hand, one should not be prepared to attribute *any* collection of false beliefs to the agent: for then there is no limit to the hopeless goals one can consistently attribute to him. Thus, effective goal-attribution demands that one use a goal-indexing mechanism which is sensitive in some way to one's beliefs about the satisfiability of the various preconditions of the relevant goals.

Yet the suggestion contained in the indexing thesis that the same goal-indexing mechanisms are responsible for *both* one's ability to perform sensible goal-attribution in the face of false beliefs on the part of the agent one is reasoning about, *and* one's ability to perform sensible goal-activation in the face of unrealised pre-conditions for the goals that one suggests. The question now arises as to whether goal attribution and goal activation really do call for exactly the *same* pattern of sensitivity to unrealised preconditions. Or, is it rather the case that the optimal goal-indexing strategy for goal-activation is different from that for goal-attribution? And if—as indeed seems likely—goal activation and goal attribution would *ideally* require different indexing mechanisms, is it feasible to construct a compromise between the two: something that serves adequately, if not optimally, for both tasks? And if so, is that what people actually do? That is: do human subjects exhibit the corresponding patterns of sensitivity to unrealised preconditions in both goal-activation and goal-attribution that the indexing thesis predicts?

As remarked above, the status of the indexing thesis is, at present, that of a suggestion for extending the simulation based theory of psychological inference. What is required, if this suggestion is to be fleshed out and evaluated as a psychological hypothesis, is a detailed investigation of the foregoing questions. The purpose of this paper has merely been to point the way to these investigations.

7 Summary

My point of departure in this paper was the simulation-based theory of psychological inference. According to that theory, one can reason about an agent's state of mind by imagining oneself in that agent's cognitive predicament and simulating his thought processes. The simulation thesis embodies the attractive idea that the same cognitive mechanisms that the reasoner uses for ordinary (non-psychological) inference get re-used in psychological inference. Thus, a reasoner does not need a *theory* of how psychological states interact; he can simply trade on the presumed similarity between himself and the agent he is simulating.

However, the simulation thesis suffers from a number of limitations, among them, the problem that simulation cannot generate hypotheses as to the *causes* of a given psychological state. The purpose of this paper is to propose the *indexing thesis* as a possible, partial solution to this problem. According to the indexing thesis, hypotheses about the psychological causes of another agent's present actions can sometimes be generated with the aid of cognitive mechanisms which are required for non-psychological goal-activation tasks. Although logically independent of the simulation-thesis, the indexing thesis also embodies the attractive idea that certain cognitive mechanisms that the reasoner uses for non-psychological inference (goal-activation) get re-used for psychological inference (goal-attribution). The cognitive mechanisms in question are mechanisms for indexing goals by conditions or actions that will achieve or help achieve those goals. We concluded by outlining some implications of the indexing thesis[3].

References

[1] Allen, James F.: "Recognising Intentions from Natural Language Utterances", in Brady, M. and Berwick, Robert C.: *Computational Models of Discourse*, Cambridge, MA: MIT Press (1984), pp. 107–166.

[2] Allen, James F. and Perrault, C. Raymond: "Analysing Intention in Utterances", *Artificial Intelligence*, **15**, **3** (1980), pp.143–178

[3] Craik, Kenneth: *The Nature of Explanation*, Cambridge: Cambridge University Press (1943).

[4] Creary, Lewis G: "Propositional Attitudes: Fregean Representation and Simulative Reasoning", *Proceedings, IJCAI-79*, Tokyo (1979).

[5] Fagin, Ronald and Joseph Halpern: "Belief, awareness and limited reasoning", *Artificial Intelligence* **34**, 1 (1987), pp. 34–76.

[6] Haas, Andrew R.: "A Syntactic Theory of Belief and Action", *Artificial Intelligence* **28**, 3 (1986), pp. 245–292.

[7] Hintikka, J.: *Knowledge and Belief*, Ithaca, NY: Cornell University Press (1962).

[8] Hobbes: *Leviathan*, (ed. C.B. Macpherson) Harmondsworth, Middlesex: Penguin Books (1968).

[9] Konolige, Kurt: *A Deduction Model of Belief*, London: Pitman (1986).

[10] Litman, D. and Allen, James F.: "A Plan-Based Recognition Model for Subdialogues in Conversation", *Cognitive Science* vol.11, no.2 (1987), pp. 163–200.

[11] Levesque, H.J: "A logic of implicit and explicit belief", *Proceedings, AAAI-84*, Austin, Texas (1984).

[12] Pratt, Ian: "Psychological Inference, Constitutive Rationality and Logical Closure", in *Vancouver Studies in Cognitive Science, vol. 1: Information, Language and Cognition*, Vancouver, BC: University of British Columbia Press (1990, forthcoming).

[13] Stalnaker, Robert C: *Inquiry*, Cambridge, Massachusetts: MIT Press (1984).

[14] Wilensky, R.: *Planning and Understanding*, Reading, MA, Addison Wesley (1983).

[3] I would like to thank Ivan Leudar for discussion of some of the ideas in this paper.