In Martin Davies and A. Stone (Eds), *Mental Simulation*, Blackwell, in press.

Simulative Reasoning, Commonsense Psychology and Artificial Intelligence*

John A. Barnden

Computing Research Laboratory & Computer Science Dept
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003-0001, U.S.A.

(505) 646-6235 jbarnden@nmsu.edu FAX: (505) 646-6218

Running Head: Simulative Reasoning in AI

Length: approx. 11,000 words

[*] This work was supported in part by grants IRI-9101354 and CDA-8914670 from the National Science Foundation, U.S.A.

ABSTRACT

The notion of *Simulative Reasoning* in the study of propositional attitudes within Artificial Intelligence (AI) is strongly related to the Simulation Theory of mental ascription in Philosophy. Roughly speaking, when an AI system engages in Simulative Reasoning about a target agent, it reasons *with* that agent's beliefs as temporary hypotheses of its own, thereby coming to conclusions about what the agent might conclude or might have concluded. The contrast is with non-simulative meta-reasoning, where the AI system reasons within a detailed theory *about* the agent's (conjectured) reasoning acts. The motive within AI for preferring Simulative Reasoning is that it is more convenient and efficient, because of a simplification of the representations and reasoning processes. The chapter discusses this advantage in detail. It also sketches the use of Simulative Reasoning in an AI natural language processing system, ATT-Meta, that is currently being implemented. This system is directed at the understanding of propositional attitude reports. In ATT-Meta, Simulative Reasoning is yoked to a somewhat independent set of ideas about how attitude reports should be treated. Central here are the claims that (a) speakers often employ commonsense (and largely metaphorical) models of mind in describing agents' attitudes, (b) the listener accordingly needs often to reason within the terms of such models, rather than on the basis of any objectively justifiable characterization of the mind, and (c) the commonsense models *filter* the suggestions that Simulative Reasoning comes up with concerning target agents' reasoning conclusions. There is a yet tighter connection between the commonsense models and the Simulative Reasoning. It turns out that Simulative Reasoning can be rationally reconstructed in terms of a more general type of reasoning about the possibly-counterfactual "world" that the target agent believes in, together with an assumption that that agent has a faithful representation of the world. In the ATT-Meta approach, the reasoner adopts that assumption when it views the target agent through a particular commonsense model (called IDEAS-AS-MODELS).

1. INTRODUCTION

The notion of Simulative Reasoning has regularly cropped up in Artificial Intelligence (Moore 1973, Creary 1979, Haas 1986, Barnden 1990, Barnden, 1992b, Ballim & Wilks 1991, Dinsmore 1991, Chalupsky 1992). It is the idea of using an AI system's own inferencing from premises P_i to a conclusion Q as a simulation of a target agent's potential inferencing from its beliefs P_i to the possible new belief Q.[1] In effect, the system temporarily adopts the premises as hypotheses, sees that the conclusion follows, and then (defeasibly) ascribes the conclusion to the target agent. The notion of Simulative Reasoning will be made somewhat more precise below. For now, the important point is that it contrasts with *non-simulative meta-reasoning*, the approach of reasoning about an agent's reasoning by means of an explicit theory *about* reasoning.

Simulative Reasoning has an obvious similarity to the Simulation Theory of mental ascription and behavior prediction as propounded by Gordon, Goldman and Harris in their chapters in this volume. However, Philosophy and AI have had rather different, though overlapping, sets of motivations for adopting the idea of mental simulation, and have focussed on different aspects of it. AI's attention has been largely confined to target agents' deductions from their *beliefs*, whereas philosophers have paid a considerable amount of attention to mental states other than belief and to the link to behavior. On the other hand, AI has developed the formal-logical and computational details of mental simulation much further, and has developed some quite sophisticated Simulative Reasoning systems (see Dinsmore, 1991, and Chalupsky, 1992, for recent cases). Philosophers have deployed the Simulation Theory partly in order to reach philosophical goals that have not featured in the relevant AI research, such as the goal of attacking functionalist approaches (Goldman, "Interpretation Psychologized," this volume). By contrast, the motivation in AI has been essentially the practical one of convenience and efficiency of computation.

Of course, the efficiency issue is of importance to Philosophy, and also to Psychology. Therefore, in the first half of this chapter I discuss the efficiency advantages of Simulative Reasoning. My hope is that by relating the Simulation Theory to AI's Simulative Reasoning I will provide a firmer basis on which philosophers and psychologists can judge exactly what computational mechanisms the Simulation Theory involves or could involve, and exactly how it differs computationally from the Theory-Theory. Indeed, one thing that will emerge is that one needs to consider the

[1] I will use "it" as the pronoun for an agent, be it human, animal, ethereal or artefactual. By a "target agent" I mean any agent that a reasoning system is reasoning about.

different varieties of simulative and non-simulative meta-reasoning carefully, as some varieties of non-simulative reasoning are not that different from some varieties of Simulative Reasoning.

The other half of the chapter yokes Simulative Reasoning to a different and rather radical proposal about how an AI system is to reason about agents' propositional attitudes. This proposal takes as central the commonsensical models of mind that are frequently used in real natural language discourse for the purpose of talking about people's propositional attitudes. For instance, speakers often talk about ideas as if they were physical objects at different physical locations in the mind conceived of as a physical container. This commonsensical view is the well-known MIND-AS-CONTAINER metaphor.[2] It is evidenced in sentences such as "Yolanda put the idea into Xavier's mind" and "Xavier hadn't brought the ideas together in his mind." There are many other commonsensical views of mind, mostly if not entirely metaphorical. For instance, it is very frequent for an occurrent, conscious thought to be portrayed as an event of internal speech, as in " 'Mike's a stupid idiot,' Xavier thought." The important point point for the present article is that these commonsense views are not used just for decoration, but are often crucial in guiding the listener in ascribing reasoning acts and beliefs to the target agents (e.g. Xavier in the examples). If Xavier hasn't brought certain ideas together in his mind, we can surmise that he hasn't drawn certain conclusions from them. If Xavier is portrayed as engaging in internal speech, then we can surmise that his thought is conscious and is operative in whatever he is doing at the moment.[3]

One major link to Simulative Reasoning in my approach is as follows. Simulative Reasoning comes up merely with suggestions about what an agent might conclude. The reasoner's answer to the question of whether the agent actually adopts such a conclusion can be greatly influenced by the commonsense models that the reasoner happens to be using as a way of viewing that agent. Although this yoking together of Simulative Reasoning with commonsense models fits with the emphasis on the defeasibility of simulative reasoning in such AI works as Creary (1979), Dinsmore (1991) and Chalupsky (1992), to my knowledge no-one in AI has proposed that rich commonsense models of mind act explicitly as the framework for that defeasibility. Also, the approach corrects the great mathematical "dryness" of almost all technically-detailed work in Philosophy and AI

[2] See, e.g., (Lakoff, 1987: p.450), (Lakoff, Espenson & Goldberg, 1989: p.61). A version of it is also one of the two metaphors that Wellman (1990: pp.268–271) identifies as central in the development of the child's theory of mind, and it plays an important role in the psychological study of idioms in Gibbs & O'Brien (1990). There is a strong connection to metaphors of communication, especially the CONDUIT metaphor (Reddy, 1979).

[3] The metaphorical quality of commonsense models of mind is extremely important, but is not the focus of the present article. My concentration is rather on their commonsensicality. I say more about metaphor in Barnden (1989), Barnden (1991), and Barnden (1992b).

on propositional attitudes. This work almost entirely ignores details of the multifarious ways real people speak and seem to think about minds. My approach can be seen as an attempt to "psychologize" the study of propositional attitudes (cf. Goldman, 1991; Goldman, "Interpretation Psychologized," this volume).

I should emphasize here that my main interest is with how AI systems (and, by presumptive extension, people) might reason about agents mental states on the basis of information coming in through natural language input. I do not presume to comment on what propositional attitudes really are, even though that is a major philosophical concern, because I do not regard it as particularly relevant to how people talk about minds or how listeners are to interpret such talk. (An analogy: the detailed chemistry of an orange is unknown to most people — even though they may know a few gross facts about substances within it — and is therefore irrelevant to mundane talk about oranges.) Therefore, my claims about psychologization are not to be assumed to have implications for the objective nature of propositional attitudes. Nor, despite my focus on natural language, do the claims directly address the question of objective, formal semantic accounts of natural language.

The concentration on natural language in my current work is largely for historical and methodological reasons. The ideas on commonsense models have their origins in research on belief representation/reasoning generally, whether or not linked to natural language understanding (Barnden, 1986). I regard the commonsense models as fundamental to how people think about people, and think that their manifestations in natural language are ultimately a secondary matter. This accords with my general agreement with emphasis of Lakoff and Johnson (1980) on the cognitive as opposed to linguistic foundations of metaphor. However, natural language discourse provides a wealth of evidence for what models are used and how they work, and also provides guidance as to what specific types of inference about belief are most worth tackling. The types of inference about belief I concentrate on are those that quite clearly need to be made by an understander in order to build a coherent view of what the discourse is conveying. However, the fundamental principles in the work could be applied also to reasoning about beliefs that are made apparent to the reasoner by means other than natural language.

The plan of the chapter is as follows. Section 2 makes the notions of non-simulative meta-reasoning and Simulative Reasoning in AI more precise, and summarises some advantages, especially gains in efficiency, resulting from Simulative Reasoning. The section also establish contact with points made in the chapters by Goldman, Gordon, and Stich and Nichols. Section 3 sketches the combination of Simulative Reasoning and commonsense-model-based reasoning that is being implemented in an AI system called ATT-Meta. The section culminates by showing that Simulative Reasoning can be rationally reconstructed as a derived view based on moregenl considerations, involving a particular, widespread, commonsense model of the mind. The reconstruction makes contact with an observation about counterfactuals made in Harris's chapter. Section 4 is the conclusion.

2. ADVANTAGES OF SIMULATIVE REASONING

The efficiency advantages of Simulative Reasoning are the main topic of this section. I make no claim that the observations are original — related, though relatively vague and brief, observations have been made by other researchers. Most of the points I make are explications of comments by Simulation Theorists such as Goldman and Gordon or by AI researchers such as Ballim and Wilks (1991), Chalupsky (1992), Moore (1973), Creary (1979), Dinsmore (1991: e.g., pp. 195ff, 214, 239), and Haas (1986).

2.1: The Task

Suppose an AI system takes a target agent, Xavier, to believe that Mike is tall, that if Mike is tall then Sue is short, and that if Sue is short then Sue is clever. More precisely, suppose the system has the following logical formulae in its own "belief" set (or "knowledge base"):

(2.1) `bel(Xavier, tall(Mike))`

(2.2a) `bel(Xavier, tall(Mike)` $\Rightarrow$ `short(Sue))`.

(2.2b) `bel(Xavier, short(Sue)` $\Rightarrow$ `clever(Sue))`.

Modal operator `bel` corresponds to the verb "believes." I stress that `bel` encapsulates a notion of "explicit" belief. That is, the belief is a piece of information that is currently available for use in Xavier's reasoning processes, rather than being an implicit belief, in other words a belief that is merely a consequence of Xavier's explicit beliefs. (Explicit beliefs need not be conscious ones, however.) I use modal logic (see, e.g., Chellas, 1980) in this article merely for simplicity of exposition. This choice of representational style makes little difference to the discussion, and in particular should not be taken to imply any commitment to special axioms or inference rules that are commonly to be found in modal logics of belief.

Our question is how the system is to come to the following reasonable though defeasible conclusions:

(2.3a) `bel(Xavier, short(Sue))`

(2.3b) `bel(Xavier, clever(Sue)).`

That is, how would the system get the effect of reasoning that Xavier performs Modus Ponens steps on his beliefs? Recall that the Modus Ponens rule is as follows:

(2.4)

$$\frac{X \qquad X \Rightarrow Y}{Y}$$

where X and Y are any logical formulae. For simplicity, we will keep to cases where the putative reasoning chain inside Xavier consists entirely of Modus Ponens steps. The discussion can readily be generalized to take account of other inference rules.

2.2: Simulative Reasoning

The Simulative Reasoning approach to concluding that Xavier believes that Sue is short is as follows. The system goes into a computational context, analogous to a procedure call, in which the following statements are temporarily adopted as facts (axioms), by stripping off the `bel(Xavier, ...)` layers from (2.1) and (2.2a):

(2.5) `tall(Mike)`

(2.6a) `tall(Mike)` ⇒ `short(Sue).`

Then, the system does a Modus Ponens application:

(2.7a)

$$\frac{\texttt{tall(Mike)} \qquad \texttt{tall(Mike)} \Rightarrow \texttt{short(Sue)}}{\texttt{short(Sue)}}$$

The result is the formula `short(Sue)`. The system now leaves the special context, and adds a `bel(Xavier, ...)` layer to `short(Sue)`, thus getting (2.3a).

In order to to conclude that Xavier believes Sue to be clever, the Simulative Reasoning process is as follows: The following are temporarily adopted as axioms by stripping of the `bel(Xavier, ...)` layers from (2.1), (2.2a) and (2.2b):

(2.5) `tall(Mike)`

(2.6a) `tall(Mike)` $\Rightarrow$ `short(Sue)`

(2.6b) `short(Sue)` $\Rightarrow$ `clever(Sue)`.

Then, the system does the Modus Ponens application (2.4a), followed by another one:

(2.7b)

$$\frac{\begin{array}{l}\texttt{short(Sue)}\\ \texttt{short(Sue)} \Rightarrow \texttt{clever(Sue)}\end{array}}{\texttt{clever(Sue)}}$$

The result is the formula `clever(Sue)`. The system now leaves the special context, and adds a `bel(Xavier, ...)` layer to `clever(Sue)`.

The crucial point is that the work done is just the work needed for the reasoner to conclude on its own account that (2.3b) holds from the premises (2.5), (2.6a) and (2.6b), plus the work needed to strip off the belief layer from each relevant belief of Xavier's, and restore one at the end. This point is of course central in discussions of the Simulation Theory, and it appears for instance in the portrayal of Simulation Theory in Stich and Nicols (this volume). We can add the observation that the effort overhead of doing the stripping and restoring is the less the more inference steps that occur in between. Of course, inference rules other than Modus Ponens may in principle be used within the simulation, and these other rules may effect styles of plausible reasoning such as induction and abduction. However, Simulative Reasoning in AI has focused on deduction by traditional inference rules.

2.3: One Alternative to Simulative Reasoning

The alternatives to Simulative Reasoning we will discuss have a Theory-Theory flavour, although as we will see the distinction between them non-simulative meta-reasoning (as in the Theory-Theory) and the Simulation Theory (or Simulative Reasoning) is murkier than it is made out to be, especially as discussions of the distinctions rarely spell out the processes in formal detail, and there are several different styles of non-simulative meta-reasoning.

Probably the simplest possible alternative to Simulative Reasoning is to appeal to applications of the following inference rule:

(2.4′) Lifted Modus Ponens

$$\frac{\texttt{bel(}A\texttt{, }P\texttt{)} \qquad \texttt{bel(}A\texttt{, }P \Rightarrow Q\texttt{)}}{\texttt{bel(}A\texttt{, }Q\texttt{)}}$$

This rule can be applied to (2.1) and (2.2a) in order to derive (2.3a), and to (2.3a) and (2.2b) to derive (2.3b). Essentially, the rule is Modus Ponens "lifted" up through one belief level. (In practice one would also need Lifted versions of inference rules other than Modus Ponens.) Clearly, the method has a significant overhead compared to just doing the base-level Modus Ponens steps (2.7a) and (2.7b). The overhead is the cost involved in continually manipulating the `bel(`A`,...)` layers in the formulae involved, rather than just manipulating P and Q.

The *continual* manipulation of explicit **bel** layers qualifies the method as being more "explicit" than Simulative Reasoning. Inference rule (2.4′) could be construed as a rule that could appear in a Theory-Theory theory of how agents A reason. At this point we come into contact with the distinction of Stich and Nichols (this volume) between the narrow and broad construals of the word "theory." A logical inference rule should be thought of as a *specification of an operation* rather than as a *statement* about something (e.g., about a agent A). Indeed, an inference rule might well, in an AI system, be *procedurally* realized as a piece of program that takes premises of the prescribed form as inputs and outputs the conclusion. The readiest example of this is the resolution inference procedure commonly used in AI reasoning systems. It is certainly not the case that rule (2.4′) need in any normal sense be explicit, and explicitly interpreted, in the system.

Therefore, in saying that (2.4′) brings us into the realm of Theory-Theory relies on a broad construal of "theory" under which the theory's nomological generalizations are merely procedurally realized.

Stich and Nichols (this volume) mention connectionism as one medium within which a broad-style theory might be couched. They do not spell out exactly how connectionist techniques would be deployed, but let us assume that there would be a connectionist network CN that took connectionist encodings of propositions of form `bel(`A`,`P`)` and `bel(`A`, `$P \Rightarrow Q$`)` as inputs, and outputed a connectionist encoding of `bel(`A`,`Q`)`. However, the move does not dispel our complaint concerning the belief layers in Lifted Modus Ponens. The type of transformation that we are asking CN to do is one that involves systematic, structure-sensitive manipulations of structures P, $P \Rightarrow Q$ and Q *embedded* within belief layers. Such embedded transformations have not yet shown to be possible in connectionist systems other than by having the system be a "mere" implementation of traditional symbolic processing. Non-implementational connectionist systems have so far merely shown some *promise* for doing systematic structure-sensitive transformations of structures that are *not* embedded in any significant sense. (See Blank, Meeden and Marshall, 1992, and Chalmers, 1990, for studies that demonstrate this promise. I comment further on the embedding issue in Barnden, 1992a.) But, *if* CN must be such an implementation, introducing it will turn out to have no philosophical advantage for Stich and Nichols.

The overhead of manipulating the belief layers in (2.4′) will be the greater the more complex the expression A is, especially as applications of the rule require that the A slot be filled consistently throughout the rule by the same agent-denoting term. Although in our example A is just the constant symbol `Xavier`, it could have been a complex description of an agent. Of course, the extra consistency-enforcing work arising in such a case could be avoided by inventing a new constant symbol to replace that description, but this is an ad hoc measure, and requires the system to work out that the step should be taken. We should observe that Dinsmore's "parochial reasoning" uses rules somewhat like (2.4′), but with the belief layers replaced by "space tags." Since these are atomic items, no matter how many belief layers are in play, the carrying-through of the space tags at every reasoning step introduces little overhead. It is therefore fair to say that the reasoning is more like Simulative Reasoning as portrayed here than it is like the method based on (2.4′).

2.4: A Further Alternative to Simulative Reasoning

Here we look at a method for reasoning about belief that could form part of a Theory-Theory account, under a narrower, more classical notion of theory.

Inference rule (2.4′) is effectively a defeasible procedural embodiment of part of the purpose of a particular axiom schema that is common in modal logics (see, e.g., Chellas, 1980; or see the Consequential Closure rule in Davis, 1990):

(2.8) $\texttt{bel}(A,\ P) \wedge \texttt{bel}(A,\ P \Rightarrow Q) \Rightarrow \texttt{bel}(A,\ Q).$

This material implication is a schema for generating a *explicit statements about* particular agents A and their beliefs. (It is far too strong, because it has lost the defeasibility of (2.4′). It forces agent A to believe all Modus Ponens consequences of its beliefs. One way of softening it would be to surround the consequent with a modal operator such as `probably(...)`. However, we will continue with (2.8) as it stands for simplicity of illustration.)

Suppose (2.1) and (2.2a) have been established. Let us assume that the following can then be inferred by Comjunction Introduction:

$\texttt{bel(Xavier, tall(Mike))} \wedge \texttt{bel(Xavier, tall(Mike)} \Rightarrow \texttt{short(Sue)).}$

Then an application of the *Modus Ponens* inference rule to this conjunction and (2.8) will establish (2.3a).

The use of (2.8) is therefore more expensive than the use of (2.4′). An application of (2.4′) has been replaced by the work of instantiating (2.8) and doing the Conjunction Introduction and Modus Ponens steps just described. This major increase in expense is is a direct result of the increase in explicitness involved in going from an implicit, procedural couching of belief consequence in (2.4′) to the explicit, declarative couching in (2.8).

It will help to clarify the increase in expense if we replace (2.8) by the following equivalent schema:

(2.8′) $\texttt{bel}(A,\ P) \Rightarrow (\texttt{bel}(A,\ P \Rightarrow Q) \Rightarrow \texttt{bel}(A,\ Q)).$

This makes little difference to the cost of the method, but it makes it more uniform my replacing the Conjunction Introduction by another Modus Ponens step. The use of (2.8′) is as follows:

- Do a Modus Ponens step on the given `bel(Xavier, tall(Mike)` and an instance of (2.8′), to get:

  $\texttt{bel(Xavier, tall(Mike)} \Rightarrow \texttt{short(Sue))} \Rightarrow \texttt{bel(Xavier, short(Sue)).}$

- Do a Modus Ponens step on the formula just derived and on the given

  `bel(Xavier, tall(Mike) ⇒ short(Sue))`,

  to get the conclusion `bel(Xavier, short(Sue))`.

Thus, the *single* application of the Lifted Modus Ponens rule (2.4′) has been replaced by *two* Modus Ponens steps. Each of these steps involves the same handling of a `bel(Xavier, ...)` layer as is required by the use of (2.4′), so the two Modus Ponens steps are certainly more expensive than the single Lifted-Modus-Ponens step. To make matters worse, the second Modus Ponens step involves considerably more complex matching of subexpressions than the Lifted-Modus-Ponens step does.

As a corollary, compared to Simulative Reasoning the current method not only involves much more in the way of belief-layer manipulation, but also replaces each ordinary Modus Ponens step within the simulation by two Modus Ponens steps of at least equal cost.

Clearly, a way of now deriving `bel(Xavier, clever(Sue))` is to go through a similar process again, using a different instance of (2.8′). In general, what happens is that each Lifted-Modus-Ponens step that would be used in the method of the previous subsection is replaced by two Modus Ponens steps, each of which is at least as expensive as that Lifted-Modus-Ponens step.

(2.8) is, of course, just one axiom schema that might appear in a particular "narrow-style" theory of agents' reasoning. Also, one might have axiom schemata concerning specific agents rather than agents in general.

2.5: Avoiding the Repeated Use of (2.8)

In subsections 2.3 and 2.4, part of our complaint was based on the repeated use of (2.4′), (2.8) or (2.8′). Each individual use involves the explicit manipulation of belief layers, so that a multisstep reasoning process will involve much more belief layer manipulation than the constant, small amount required by simulative reasoning. So, can we avoid the repeated application?

A reasoner with standard deductive capabilities can prove the following:

(2.9)

`[tall(Mike) ∧ (tall(Mike) ⇒ short(Sue)) ∧ (short(Sue) ⇒ clever(Sue))]`

`⇒ clever(Sue).`

This is a tautology, and merely encapsulates some logical relationships. It has nothing to do with whether Mike is actually tall, and so on. Let us assume for the sake of argument that the reasoner can defeasibly ascribe to any agent A a belief in any tautology. (This reflects a common inference rule in modal logics of belief.) So the reasoner can suppose that Xavier believes (2.9). Let us suppose that the reasoner can infer that Xavier believes the antecedent of (2.9), because he believes each conjunct. So, a *single* use of (2.4′) or of (2.8) suffices now to establish the consequent of (2.9), in other words (2.3b). In this single use, P and Q are the antecedent and consequent of (2.9), respectively.

Although this method does eliminate the repeated usage that stood accused, it still has considerable overhead compared to Simulative Reasoning. For one thing, there is the need to do the conjunction introduction, along with the entailed manipulation of belief layers (those in the statements (2.1), (2.2a) and (2.2b)). This is extra work compared to Simulative Reasoning, because the latter's stripping off of belief layers is already paralleled by the similar stripping off needed to prepare for the above process of proving (2.9). And there is also the need to finish the job by using (2.4′) or (2.8) (and the necessary Modus Ponens step in the latter case).

2.6: Nested Belief

In AI it is commonplace to apply simulative reasoning to nested beliefs (Ballim & Wilks, 1991; Chalupsky, 1992; Dinsmore, 1991). For instance, if the system knows that

Yolanda believes that Xavier believes that Mike is tall

Yolanda believes that Xavier believes that if Mike is tall then Susan is short

Yolanda believes that Xavier believes that if Susan is short then Susan is clever

then it is reasonable for the system to conclude, defeasibly as always, that Yolanda believes that Xavier believes that Susan is short, and that Yolanda believes that Xavier believes that Susan is clever. Simulative Reasoning can be applied to such nested cases much as before, the difference being that the system is not adopting beliefs it presumes Xavier to have, as input hypotheses for the simulation, but is rather adopting beliefs it presumes *Yolanda to believe* Xavier to have. The extra overhead introduced by the Yolanda layer is therefore just the need for an increased amount of stripping off and restoration of belief layers at the start and end of simulation.

The advantage of Simulative Reasoning over methods such as those of sections 2.3–2.5 is magnified by every extra level of belief nesting. In the case of the method of section 2.4, the single layer of belief caused the amount of work to be doubled at least, because each Modus Ponens step in Simulative Reasoning is replaced by two Modus Ponens steps of at least equal cost. But adding the extra Yolanda layer then *further* doubles the amount of work compared to Simulative Reasoning, and so on similarly for each extra belief layer. This observation is a special case of one made by Haas (1986).

It is interesting to consider, however, whether the less explicit method of section 2.3 suffers so drastically under nesting. Recall that this method uses Lifted Modus Ponens, and our complaint was to do with the amount of belief-layer manipulation required, not with any amplification of the number of reasoning steps. The trouble is that the Lifted Modus Ponens rule (2.4′) is not able by itself to deal with nesting. Consider the task of inferring that

Yolanda believes that Xavier believes that Susan is short.

The system would need to perform an application of Lifted Modus Ponens in which

A is `Yolanda`

P is `bel(Xavier, tall(Mike))`

Q is `bel(Xavier, short(Sue))`.

But this application requires the system first to establish

`bel(Yolanda, bel(Xavier, tall(Mike))⇒ bel(Xavier, short(Sue)))`.

Note that this is different from the premise that Yolanda believes that Xavier believes that if Mike is tall then Susan is short, i.e.:

`bel(Yolanda, bel(Xavier, tall(Mike) ⇒ short(Sue)))`.

Hence, other inference rules or axioms would have to be brought into play to enable the desired application of Lifted Modus Ponens. In fact, the cheapest way to modify the Lifted Modus Ponens method to allow for an extra level of belief nesting is probably propose the following extra rule:

(2.4″) <u>Doubly Lifted Modus Ponens</u>

$$\frac{\begin{array}{l}\texttt{bel}(A_2,\ \texttt{bel}(A_1,\ P))\\ \texttt{bel}(A_2,\ \texttt{bel}(A_1,\ P \Rightarrow Q))\end{array}}{\texttt{bel}(A_2,\ \texttt{bel}(A_1,\ Q))}$$

Clearly, yet another rule would be needed to account for a further level of nesting, and so on. The resulting sequence of rules is ever-increasing in the belief-layer manipulation overhead it entails.

2.7: Other Considerations

As I mentioned in the Introduction, AI has concentrated on the ascription of its own *straightforward deductions* to other agents. However, Simulative Reasoning is obviously able to apply without modification to other forms of reasoning, such as abduction, induction, default reasoning, analogy-based reasoning, or whatever. The more types of reasoning there are, the more extra axioms or inference rules are needed by a non-simulative meta-reasoning system (by analogy with the extra axioms or rules in sections 2.3–2.5).

In AI and Psychology one needs to be very conscious of the *control* issue in inference. Outside these fields, there is a tendency to concentrate on what axioms and inference rules are used, at the expense of the question of *when* they are used and *why.* But if one takes control issue seriously, one realizes that the AI system not only has to have adequate control over its own inferences, but must also be able to ascribe reasoning control (adequate or otherwise) to other agents, unless for some reason it were to be sufficient to take a given target agent to make inferences in a totally indeterminate way. Now, it seems reasonable, as a default, for the system to take the target agent's control regime to be the same as its own The claim now is that Simulative Reasoning allows this default to be adopted much more efficiently than non-simulative meta-reasoning does. This is because whatever control heuristics are used in the course of the Simulative Reasoning are *ipso facto* ascribed to the agent in question. For instance, suppose the control heuristics have the effect that the Simulative Reasoning episode comes up with conclusion Q as opposed to some other possible conclusion Q'. Then this choice will be automatically ascribed to the agent whose reasoning is in question.

To get this effect in non-simulative meta-reasoning, two strategies are available. In one, the system's meta-reasoning itself would be governed by control heuristics that are designed to parallel the base-level control. For instance, in our Q/Q' example, the meta-level control is designed to have the effect that the Q-relevant instances of, say, (2.4$'$) are concentrated on at the expense of the Q'-relevant instances. The second strategy is for the system to have explicit facts *about* the agent's control. These facts would explicitly state that under such-and-such conditions, such-and-such a choice of proof-path (at the base level) is made. In either case, major extra machinery is needed. Also, the control issue further magnifies the increased cost of the method of section 2.4

in comparison to Simulative Reasoning. We saw that that method doubled the number of Modus Ponens steps every time a belief layer is added. But the longer the reasoning sequences, the greater the control problem. This point is made by Dinsmore (1991: p.85).

An additional, potential motivation for adopting Simulative Reasoning is that it is a very natural extension of the subproof technique in "natural deduction" methods. This point plays an important role in Chalupsky (1992) and Dinsmore (1991). A subproof is a region marked off within a larger (sub)proof. One type of subproof is used to establish a conclusion of form $R \Rightarrow S$. The formula R occurs as a hypothesis at the start of the subproof. Reasoning steps of any sort and number are used to derive S. The subproof is then exited, and the hypothesis "discharged" by converting S to $R \Rightarrow S$. If, therefore, a system already uses the subproof technique for quite ordinary deductive purposes, it is only natural to adopt Simulative Reasoning as the method for reasoning about beliefs. The simulation is just what goes on inside a new type of subproof, with the premise beliefs acting as the initial hypotheses in the subproof.

Simulative Reasoning has an efficiency advantage distinctly different from those so far presented. A cognitive system that learns to operate in the world, rather than being fully preprogrammed, may need to *adjust its reasoning mechanisms.* The system may well have to learn the appropriate conditions for the application of various styles of reasoning, the detailed control of them, the precise effects they should have (in the non-deductive styles), and so forth. This applies just as much to reasoning that the system ascribes to other agents as to its own reasoning. The Simulative Reasoning approach has the great advantage that adjustments only need be made once, to the rules, axioms or whatever that the system uses for ordinary reasoning that does not touch on agents' beliefs. The adjustments are then in effect automatically ascribed to other agents, because of the changed nature of the simulation of those agents. By contrast, non-simulative methods of belief reasoning would require separate adjustments to the axioms or inference rules, such as (2.4′) or (2.8), for handling belief.

Finally, Simulative Reasoning has a technical advantage that appears not to have been previously noted. It is clear that any sort of reasoning about agents' reasoning must be defeasible. In the non-simulative methods described above, this defeasibility must be encapsulated in the individual axiom schemata and rules of inference such as (2.4′) or (2.8), as we pointed out. This adds extra complications to the syntax of such entities and/or to the control mechanism whereby they are used, and the required style of defeasibility is distributed across multiple axiom schemata and rules. On the other hand, in Simulative Reasoning the simulations can use ordinary, *non*-defeasible Modus Ponens (and similarly for other traditional inference rules), because the reasoner should only defeasibly take the outputs of the simulation to be believed by the target agent. Thus, there is a greater localization or modularization of the defeasibility management mechanisms, making the system simpler overall and more easily adjustable in the light of experience. Also, the efficiency advantage of Simulative Reasoning is further amplified by relieving the innards of the simulation from the responsibility of handling the defeasibility.

The previous paragraph was only about the reasoner's defeasible reasoning about the target agent's *non*-defeasible reasoning. In particular, we referred to the modelling of the target agent's Modus Ponens steps. The point is that *if* the target agent takes a Modus Ponens step then it will be non-defeasible — but the reasoner cannot be sure that the agent *will* take the step. What about defeasibility within the target agent? In that case we have to deal with defeasibility at two distinct levels. The defeasibility at the lower level (i.e., within the target agent) *will* appear within a simulation by the reasoner. For instance, if the target agent is taken to perform induction, then inductive steps will occur within the simulation. So, the type of defeasibility that is removed from within simulations is only the higher-level type, concerned with the *reasoner*'s uncertainty concerning whether the target agent has actually performed the reasoning that has been simulated.

2.8: Some Connections to Other Chapters

Goldman, in his "Empathy" and "Interpretation Psychologized" chapters, points out that the Simulation Theory absolves the reasoner from knowing the nomological properties of decision making, and, in particular, absolves children from having to learn elaborate, arcane rules about people's mental processes. Note that this latter point is related to but distinct from the observation in the previous subsection about the ease of adjusting reasoning mechanisms. In claiming this ease I do not claim that the adjustment process that would be needed in a Theory-Theory account would be questionable in principle — merely disadvantageous in practice. Goldman's point about explicit knowledge of rules of thinking and about children's learning of them is more a matter of principle.

Gordon objects to the idea that the Simulation Theory involves the reasoner in *using itself as a model* for the agent reasoned about. One has to be careful here. According to the Simulation Theory and Simulative Reasoning the reasoner does *serve* as a model of the target agent, and in that sense does "use itself" as a model. This is the reason why "simulation" is an appropriate term. But this does not mean that the reasoner *explicitly regards itself* as a model of the agent. It would be possible in principle for the reasoner to lack the meta-cognitive capabilities that would be needed for it to realize that it was undertaking simulation of other agents. The agent just does the simulation, and need not have *any* knowledge of or beliefs about how the conclusions produced by simulation arise. Therefore, I simulate Gordon as meaning that the Simulation Theory does not require an agent to know that it is doing simulation.

One advantage for spelling out the nature of Simulative Reasoning in clear computational terms is that it becomes obvious that there is no necessary involvement of propositions like "I am simulating agent A" or "My reasoning processes are acting as a model of A." The simulating reasoner must, certainly, keep track of the fact that conclusions it comes up with are to be attributed to the particular target agent. This keeping track could perhaps be by means of explicit propositions akin to "My conclusions are really Xavier's."[4] But note that the reasoner does not need to know that those conclusions are being produced by *simulation.* As far as the reasoner is concerned, and insofar as the reasoner maintains propositions about its own reasoning activity, it can just view itself as reasoning, *tout court,* about Xavier. This is not to say that the reasoner should not, or that human reasoners do not, view themselves as simulating. I am merely supporting Gordon to the extent of saying that such a view is not a necessary feature of the Simulation Theory.

It seems it me that the issue of knowledge about one's own simulating has been obscured by the use of examples in which the reasoner is *consciously* thinking about another agent. Since consciousness already has a self-reflective quality, it is easy to abstract from such examples the idea that agents are aware of their (putatively) simulative nature. Unfortunately, the sheer giving of examples introduces a bias towards conscious ones, because the reader is implicitly invited to *consciously* imagine being the reasoner thinking about some target agent, and it is then natural for the reader to imagine conscious thinking. Here I resist the tendency, apparent for instance in Harris (this volume, see his concluding section and rebuttal of Stich and Nichols's belief-perseverance argument), to cast the reasoner's reasoning module, used to simulate other agents, as being concerned particularly with conscious reasoning or decision making. (In fact, removing the consciousness attribute does not hurt Harris's rebuttal.) My stance springs no doubt from my AI background,

[4] An anecdote may be in order. My daughter when two or three years old used to playfully imagine crazy situations such as hippotamuses leaping onto our car, and working through the possible dire consequences. After a while she would forget that she was pretending, and become afraid that the imagined events were going to happen.

given the (regrettable) bias in AI against considering issues of consciousness versus unconsciousness. Also, AI researchers are probably much readier than, say, some psychologists to grant that complex inference processes, working on complex propositions, often proceed unconsciously. Indeed, it is hard to see how we could otherwise understand even simple natural language discourse. Of course, there is no reason to think that unconscious reasoning should not involve propositions about the reasoner itself — there is more to consciousness than self-representation, which is fairly commonplace in AI.

Stich and Nichols object to "Argument 4" for the Simulation Theory, which is the claim that the theory is much simpler than the Theory-Theory. Their objection is that claims for the simplicity have ignored the "*very* non-trivial" task of getting the "control mechanism" to work smoothly. This control mechanism is what takes the reasoner's reasoning apparatus "offline," feeds it "pretend" beliefs as inputs, and handles the output suitably. For vividness I will rename this mechanism as the "simulation cocoon." Because of the non-triviality, Stich and Nichols claim that neither side in the debate can currently claim much advantage of simplicity, and that for proper comparisons we need to wait for up-and-running systems. Well, there are implemented AI systems that use Simulative Reasoning, and the task of devising the simulation-management mechanism, while certainly non-trivial, has been performed without tremendous difficulty. It remains to be seen whether a system for belief reasoning that does not use Simulative Reasoning can be devised and made to do belief reasoning that is as complete and pragmatically realistic as that done by Simulative Reasoners. However, Stich and Nichols are certainly right to point out that one must keep the simulation cocoon firmly in mind when discussing the Simulation Theory. It plays an important role in the next subsection.

2.9: Collapsed Cocoons and the Transitivity of Simulation

In discussing nested belief in section 2.6, I implied that Simulative Reasoning would operate as follows:

(1) Strip off the "Yolanda believes that Xavier believes that" from given propositions, getting the simulation inputs

Mike is tall

if Mike is tall then Susan is short.

(2) Within the simulation, infer that Susan is short by Modus Ponens.

(3) Exit from the simulation and conclude that Yolanda believes that Xavier believes that Susan is short.

Steps (1) and (3) are actions of the simulation cocoon. The propositions listed in (1) are propositions that Xavier believes *according to Yolanda* and the simulation in (2) is a simulation of Xavier's reasoning *as viewed by Yolanda.*

However, the process (1) to (3) needs some justification, as it does not really drop straight out of the simulation approach. After all, the system is meant to be reasoning about Yolanda's reasoning, so to conform to the notion of Simulative Reasoning *in general* what the system should be doing is a simulation in which the inputs and outputs are propositions that *Yolanda* believes and in which the reasoning actions are meant it simulate *Yolanda's* reasoning actions. It just so happens that Yolanda's beliefs and reasoning are about Xavier's beliefs; but that is just a special case. Of course, it is natural (in the absence of contrary evidence) for the reasoner to take Yolanda herself to be reasoning about Xavier *by means of simulation.* So, we have one simulation nested within another:

(1′) Strip off the "Yolanda believes that" layer from given propositions, getting simulation inputs

Xavier believes that Mike is tall

Xavier believes that if Mike is tall then Susan is short.

(2′) Do the outer simulation: infer that Xavier believes that Susan is short, by means of an inner simulation.

(3′) Exit from the outer simulation and conclude that Yolanda believes that Xavier believes that Susan is short.

Here (2′) is expanded into:

(2′) Within the outer simulation, do the following:

(2′/1) Strip off the "Xavier believes that" layer (from the two propositions listed in (1′), getting inner simulation inputs:

Mike is tall

if Mike is tall then Susan is short.

(2′/2) Do the inner simulation: infer that Susan is short by Modus Ponens.

(2′/3) Exit from the inner simulation with the conclusion that Xavier believes that Susan is short.

This inner simulation, (2′/2), is to be interpreted as a simulation of Yolanda's (presumed) simulation of Xavier's reasoning. We also have (2′/1) and (2′/3) as simulated actions of *Yolanda's* (presumed) simulation cocoon. Therefore, the "official" version of how Simulative Reasoning should work in nested cases involves not one simulation cocoon as in the process (1) to (3) but two. However, step (2′/1) immediately follows step (1′), and (2′/3) immediately precedes (3′), provided that Yolanda's reasoning is *only* about Xavier's beliefs. In this special case, therefore, the system can streamline the process by collapsing the cocoons together, combining (2′/1) with (1′) and (2′/3) with (3′). The system can therefore just perform (1) to (3). Thus, the original process we started with is vindicated, but only as an optimized special case. If Yolanda is meant to be reasoning about things other than Xavier's beliefs, or has to work out that Xavier has a particular belief from some proposition like "Xavier is a communist," then reasoning actions may be interposed between (2′/1) and (1′) and/or between (2′/3) and (3′), forcing double cocoonery. (In fact, this more elaborate method appears to be the one performed in all cases by the system of Chalupsky, 1992.) I see this complication as supporting Stich and Nichols's point (this volume) that we need to be careful to pay attention to simulation cocoons.

Crucial to the streamlined or collapsed method is the fact that a simulation of a simulation of a process X is a simulation of X: that is, simulation is *transitive*. (Analogously, a copy of a copy of a document is a copy of that document; a model of a model of an aeroplane is a model of the aeroplane.) I believe that this transitivity is the key to, or at least epitomizes, the ever increasing advantages of Simulative Reasoning as more nesting levels are considered (cf. section 2.6).

3. COMMONSENSE MODELS OF MIND

The AI system I am developing is called ATT-Meta (ATT for propositional ATTitude, Meta for Metaphor). It is directed at the interpretation of English text in which mental states are described. In particular, it is directed at seeing how different statements that have implications for mental states cohere with each other. The system is not yet fully implemented, but the representation and inferential schemes have been designed in considerable detail (Barnden, 1989, 1990; Barnden, 1992b) and are in the process of being implemented. For the purposes of this chapter, the main theses behind ATT-Meta are as follows.

(A) Speakers often think and talk about mental states in terms of a variety of different commonsense models of mind, including MIND-AS-CONTAINER, IDEAS-AS-MODELS, and IDEAS-AS-INTERNAL-UTTERANCES (to be described below).

(B) The commonsense models are not merely decorative — they are often important to the addressee's understanding. Defeasible inferences (about mental states) that can be drawn under the guidance of these models are important in perceiving discourse coherence.

(C) Those model-guided inferences should be drawn with the aid of Simulative Reasoning.

There is room for only a sketch of the justification for the theses. I have written more extensively about them elsewhere (Barnden, 1990; Barnden, 1992b). Consider the following sentence.

(3.1A) *Xavier fixed a meeting with Mary for 10am, although in one part of his mind he knew that he had arranged to meet Peter then.*

I assume for simplicity that a meeting never involves more than two people, so that (3.1A) reports a conflict of arrangements. Our focus will on the "in one part of his mind" qualification. Such reference to "parts" of a mind is common in real text, and is readily understood by speakers of English. I make the following claims based on my own intuitions and the intuitions of informants.

(a) The speaker is describing Xavier's mind on the basis of a commonsense model, namely MIND-AS-CONTAINER, that casts the mind as a physical container. Ideas (beliefs, in particular) are cast as physical objects that can be in various places in the container.

(b) A defeasible implication of the subordinate clause in (3.1A) (starting at "although") is that there was another part of Xavier's mind that *lacked* the belief about the 10am meeting with Peter.

Let us call the explicitly mentioned part of Xavier's mind the Peter part and the other part the non-Peter part.

(c) The implication in (b) provides a reasonable explanation of Xavier's action in making the Mary arrangement; *viz.,* the non-Peter part was the one operative in making the arrangement, so that in some sense Xavier had temporarily forgotten about the Peter arrangement.

I stress at once that it is by no means certain that the explanation in (c) is correct — we are well within the realm of defeasible, merely-plausible inference. One alternative possibility is that Xavier was aware of the clash, but this in itself was of lesser importance than getting the Mary arrangement fixed; and we might surmise that Xavier was at some point going to cancel the Peter arrangement. However, I suggest that the explanation (c) is of at least comparable plausibility to this one.

We can get a sharper impression of what is going on in (3.1A) by contrasting it with a modified fragment that uses a different commonsense view of mind:

(3.1B) *Thinking to himself, "I should meet Peter at 10am," Xavier arranged a meeting with Mary for that time.*

I submit that a very strong implication of this sentence is that, in the act of making the Mary arrangement, Xavier knew full well that he should meet Peter at 10am. There is now no question of the Peter belief being shoved into some corner of his mind [sic]. Rather, appeal is being made to the IDEAS-AS-INTERNAL-UTTERANCES model, which casts conscious, occurrent, dominant thoughts as *internal natural language utterances*. Thus, (3.1B) favours an explanation different from that in (c) of why Xavier made the Mary arrangement. The forgetting explanation in (c) becomes highly implausible if not downright impossible. A much more plausible explanation now is that Xavier is deliberately snubbing Peter. This had no particular support from (3.1A), though it is consistent with it.

One might query the need for the listener to come to an explanation of Xavier's Mary arrangement at all. We should therefore observe that suitable continuations of the narrative could make an explanation important. For instance, consider the following candidate continuation:

(3.1C) *The deliberate snub was just one of many that had happened.*

This fits easily with (3.1B). It is clear what "the deliberate snub" is. If the listener has already (tentatively) concluded that Xavier was deliberately snubbing Peter, then the snub reference can immediately be resolved. If the listener has not formed that conclusion, she is now nudged into doing so, and its fit with (3.1B) becomes apparent. By contrast, (3.1C) seems to be an unlikely continuation of (3.1A). It is by no means impossible: conceivably, it is only the part of Xavier that knew about the Peter arrangement that was deliberately snubbing him — i.e., in some sense Xavier was not whole-heartedly [sic] snubbing him. Another, more tenuous possibility is that the snub was actually one that has already been alluded to in the narrative before (3.1A), has nothing to do with Peter, and is to be the topic of Xavier's meeting with Mary. Overall, however, having (3.1C) as a continuation of (3.1A) raises much more of an inferential puzzle than having it as a continuation of (3.1B) does. The reason for this lies in the implications of the commonsense models. Notice that the puzzle is one that must indeed be tackled, on the assumption that the listener does try to discover the referent of "The deliberate snub" in (3.1C).

I should reinforce claim (b) above. I do so by appealing to a partial analogy with the following fragment:

(3.2) *There was a red-plush carpet in one part of the restaurant. Yolanda decided to sit somewhere else.*

To me and others, the first sentence here carries the very strong suggestion that there was another part of the restaurant that did *not* have a red-plush carpet. It is difficult to see why the "in one part ..." phrase would be included otherwise. (Moreover, the fragment suggests that Yolanda decided to sit somewhere else *because* there was no red plush carpet there.) Just as the first sentence in (3.2) has the mentioned suggestion, so the first sentence in (3.1A) implies that another part of Xavier's mind did not contain the knowledge about the Peter meeting.

A complication worth mentioning is that (b) does not of itself claim that the non-Peter part contains a belief *contrary* to the Peter-arrangement belief. I take (3.1A) to be uncommitted as to whether such a contrary belief exists or not. This uncertainty itself has an analogy in (3.2), because it is possible that the other part of the restaurant has, say, a blue carpet. Also, the implication of the first-sentence in (3.2) that another part lacks the carpet is defeasible, just as the implication in (b) is. It would be defeated if the second sentence had been, say, "Yolanda later realized that it was all over the restaurant."

I now consider certain aspects of the reasoning that ATT-Meta should perform when interpreting (3.1A). (A similar example is analyzed in much more detail and completeness in Barnden, 1992b, where also the (3.1B) case is considered.) To cut a long story short, I consider just the particular question of how ATT-Meta could defeasibly conclude that Xavier does not infer a conflict between the Peter arrangement and the Mary arrangement. More precisely, given that he has the belief *meet-Peter* that he should meet Peter at 10am, and the belief *meet-Mary*, how come he does not infer a contradiction? I will also consider at the same time the question of how ATT-Meta is to surmise that Xavier *does* notice the contradiction in the situation reported in the following alternative continuation of (3.1A):

(3.1D) *When Xavier brought the arrangements together in his mind, he hurriedly phoned both Peter and Mary.*

I claim that the "bringing together" of the two beliefs about the arrangements enabled Xavier to see the conflict, whereas beforehand he could not see it because the beliefs were in different parts of his mind. Notice that "bringing together" is a natural type of event in the MIND-AS-CONTAINER view.[5]

To oversimplify a little, ATT-Meta contains the following two defeasible-inference rules (which presuppose that agent A's mind is being viewed through MIND-AS-CONTAINER):

(3.3–)

Beliefs $P_1, ..., P_n$ are not all close together in A's mind.

Q follows from $P_1, ..., P_n$.

Q does not appear in A's mind (at present).

(3.3+)

Beliefs $P_1, ..., P_n$ are all close together in A's mind.

Q follows from $P_1, ..., P_n$.

Q appears in A's mind (a little later).

Rule (3.3–) is what we need for dealing with (3.1A), before a continuation such as (3.1D) appears. I assume that ATT-Meta realizes that the force of the subordinate clause in (3.1A) is that the focus is on the *non-Peter* part of Xavier's mind, even though this is not the explicitly mentioned one. ATT-Meta assumes that Xavier believes that he will meet Mary and (at least tentatively) that this belief (*meet-Mary*) is in the non-Peter part. The latter assumption is because it is the non-Peter part that is in focus. ATT-Meta now infers that, by virtue of being in different parts of Xavier's mind, beliefs *meet-Mary* and *meet-Peter* are not close together. ATT-Meta now uses

[5] I assume in the present example that the phrase "the arrangements" in (3.1D) is actually a *metonymic* reference to some of Xavier's *beliefs about* those arrangements. See Fauconnier (1985) for discussion of metonymies similar to this.

rule (3.3–) to infer that Q, a proposition stating the conflict between the arrangements, does not appear in Xavier's mind. We can take n to be 2, P_1 to be *meet-Peter*, and P_2 to be *meet-Mary.*

We now see where Simulative Reasoning fits in. We need to assume that ATT-Meta has seen that Q follows from P_1 and P_2. This could either be done by non-simulative meta-reasoning or Simulative Reasoning, just as in section 2, and of course I suggest that Simulative Reasoning be used. Notice that we are using the epithet "Simulative" here even though the ultimate conclusion will be that Xavier does *not* form conclusion Q.

It should be clear that the force of (3.1D) is to tell ATT-Meta that P_1 and P_2 are now close together, allowing (3.3+) to be used to infer that Q did then appear in Xavier's mind. Again, Simulative Reasoning could be used to establish the second premise, although if this has already been done for the (3.3–) application, the system could perhaps have remembered that Q followed from the P_i.

With (3.3–) and (3.3+) as they stand, the MIND-AS-CONTAINER model is alluded to by the "[not] close together" qualifications in the first premise. However, the second premise has nothing to do with MIND-AS-CONTAINER, and the rules as wholes therefore do not just drop out of that commonsense model. But there is a way of making them drop out, and this the method being implemented in ATT-Meta. It is described in Barnden (1992b).

3.1: A More Fundamental Connection to Simulation Theory

So far, Simulative Reasoning has been portrayed as a rather ancillary aspect of ATT-Meta. However, there is a sense in which it is actually a consequence of the way ATT-Meta uses commonsense models. In order to explain this, I need first to provide a rational reconstruction of the idea of Simulative Reasoning (and, by extension, of Simulation Theory).

As we saw in section 2.8, Simulative Reasoning does not enforce the notion that the reasoner itself realizes that it is *simulating* the target agent, Xavier. As far as the reasoner is concerned, it is reasoning about *part of the world that it takes Xavier to believe in,* in exactly the same way as it would reason about the "real world," namely the world the reasoner believes in. This world Xavier believes in, which we can call "Xavier's world," may or may not accord with the real world; if Xavier's beliefs are false (as far as the reasoner is concerned), it will be a counterfactual world (as far as the reasoner is concerned). Suppose now the reasoner takes, by default, Xavier to possess a *faithful representation of his world*, or at least of the part that the reasoner is thinking about;

that is, if something happens or holds in Xavier's (partial) world, then Xavier believes it does. Of course, the reasoner will never conclude, by this route, that Xavier believes something that it has not itself concluded about Xavier's world; and it will conclude that Xavier believes anything that it itself concludes about Xavier's world. Therefore, the effect is just the same as what happens under the normal construal of Simulative Reasoning. We can still use the term Simulative Reasoning without embarrassment, because the reasoner's reasoning is still a simulation of what Xavier's reasoning process would be if it were an exact copy of the reasoner's. However, we have separated the notion of Simulative Reasoning into two legs:

(A) the reasoner's own reasoning about Xavier's world;

(B) Xavier's having a faithful representation of (the relevant part of) his own world.

Given that Xavier's world is counterfactual if his beliefs are false, I find it interesting that Harris (this chapter) should imply that the ability to engage in counterfactual reasoning appears to arise at roughly the same time as the ability to reason correctly about false beliefs (see his discussion of Step 4 in the section on the Explanatory Strategy of ST). Also, as Gordon and Goldman point out in their defences of the Simulation Theory, a reasoner needs to be able to engage in counterfactual reasoning anyway, so rationally reconstructing Simulative Reasoning and the Simulation Theory as being based on a form of reasoning that includes the counterfactual type is apposite. Of course, Xavier's world need not be counterfactual, so it is wise for us to liken Xavier's world, in general, to a *story* world rather than a counterfactual one, since a story world can be true or false. Also, note that Dinsmore (1991) casts reasoning about belief as just one special case of *parochial reasoning* within "spaces," i.e. (partial) worlds, other special cases being story spaces and counterfactual spaces. Dinsmore's spaces are inspired by the "mental spaces" of Fauconnier (1985).

Now we can view the ATT-Meta approach as generalizing the view we have just set up, by allowing the reasoner to view Xavier as having any one of a large variety of different relationships to his world, some of them less than completely faithful. Thus, if the reasoner (ATT-Meta, say) views Xavier through the MIND-AS-CONTAINER metaphor, then the reasoner takes Xavier to contain ideas about the world, where those ideas may not come into contact with each other, and may therefore not lead to a conclusion that is indeed true of his world. So, Xavier may contain the idea that Mike is tall and the idea that if Mike is tall then Sally is short, but may not bring these ideas together, and may therefore *not* come to believe that Sally is short. This is the case even though the reasoner itself has concluded that Sally is short in Xavier's world.

As a *special case* of the ATT-Meta approach, Xavier can still have a completely faithful representation of his world. In fact, as explained in Barnden (1992b), ATT-Meta uses a certain *default commonsense model* through which to view agents; and this default model is one that ensures faithful representation by the agent. The default model is called IDEAS-AS-MODELS (and is a very special case of the MIND-AS-CONTAINER model). When ATT-Meta views Xavier through IDEAS-AS-MODELS, it takes Xavier's mind to *contain a "model" of his world*, or at least of the part of it that is currently relevant to ATT-Meta. This MODEL inside Xavier is a model of his world in much the analogical sense that a realistic theatrical production is a model of some actual or hypothetical world. The model contains counterparts to things in Xavier's world, and counterparts to the relationships between and events involving those things. For the purposes of this paper the crucial point is that the MODEL is a faithful representation of (part of) Xavier's world. IDEAS-AS-MODELS is, I claim, evidenced in such forms of expression as "In Xavier's mind, Mike was taller than Sue."

To put all the above together: when using its default, IDEAS-AS-MODELS view of agents ATT-Meta *therefore* engages in Simulative Reasoning of them. Notice that in a sense we have turned the direction of simulation round — it is *Xavier* that is directly portrayed as doing simulation, of his own world. (And observe that the simulation need not be simulation of *reasoning* particularly.) However, because of the faithfulness of this simulation, we can take ATT-Meta's reasoning about Xavier's world to be a simulation of Xavier's reasoning about (i.e., simulation of) his world. But Simulative Reasoning is now a derived view.

If Xavier's world happens to contain believing agents, Zoltan for instance, *and* ATT-Meta takes Xavier to take the IDEAS-AS-MODELS view of Zoltan (the default case), then it turns out easily enough that Xavier can be construed as doing Simulative Reasoning about Zoltan's reasoning, and that ATT-Meta can be construed as doing Simulative Reasoning about Xavier's Simulative Reasoning.

The use of IDEAS-AS-MODELS as a default was motivated by considerations independent of any attempt to link ATT-Meta's approach to Simulative Reasoning; indeed, I only recently saw the deep connection that has been described here. I speculate that the IDEAS-AS-MODELS commonsense model of thinking is strongly related to, if not essentially identical, to the "container-copy metaphor of mind" that Wellman (1990: p.269) discusses. If this speculation is correct, it could open up a route to integrating the ATT-Meta approach into an understanding the development of adults' reasoning about reasoning.

4. CONCLUSION

We have looked at the Simulative Reasoning idea as it appears in AI. It is a relatively convenient and efficient way of reasoning about reasoning, compared to non-simulative meta-reasoning. The motivation behind Simulative Reasoning is therefore highly practical, rather than to do with matters of philosophical principle. Nevertheless, relating the psycho-philosophical Simulation Theory to precise computational frameworks for Simulative Reasoning in AI has the benefit of clarifying exactly what that theory involves in the way of mechanisms and exactly how it compares on that score with non-simulative meta-reasoning as in the Theory-Theory. We saw in particular that non-simulative meta-reasoning has different varieties, some closer to Simulative Reasoning than others. This point is related to Stich and Nichols's observations about broad construals of the word "theory."

In section 3, Simulative Reasoning was seen only as a source of suggestions about what agents might infer. Whether these suggestions are adopted can depend on information within commonsensical (and largely metaphorical) models of mind that may happen to be in use for characterizing an agent's mental state. Thus, commonsense models can in effect be used as a filter on the suggestions of Simulative Reasoning. The general idea of filtering the suggestions appears in Creary (1979), but to my knowledge the use of commonsense models of mind in the filtering has not previously been proposed.

At the end of the section we came to a revised and more integrated view of the role of Simulative Reasoning in the approach. We viewed the reasoner's reasoning to be composed of (a) reasoning about the situations in the "world" the agent believes in, and (b) an assumption that the target agent has a representation of that world, where (c) that representation is viewed by the reasoner through any one of the large variety of commonsense models of mind. Then, Simulative Reasoning is simply the special case obtained by letting the model in (c) be the IDEAS-AS-MODELS model, one effect of which is to cast the target agent as having a faithful representation of its world. Since I propose that IDEAS-AS-MODELS be used as the *default* model, it turns out that Simulative Reasoning is the default way of reasoning about agents' reasoning. IDEAS-AS-MODELS may connect to Wellman's (1990: Ch.9) claim that the view of belief plays an important role in development. Also, since the agent's world may well be counterfactual, the approach connects to Harris's comments on counterfactuals in his chapter. The reasoning in (a) is exactly of the type that the reasoner needs to use in reasoning about counterfactual situations in general, and about the situations in stories, whether true or false. The reasoning is thus closely related to the "parochial reasoning" of Dinsmore (1991).

I view commonsense models of mind as "folk-psychological," provided that term is taken merely to refer to how the folk commonsensically view the mind. Goldman's attack (in "Interpretation Psychologized, this volume) on folk-psychological theories might lead one to think that I am in conflict with his work. This is illusory, because Goldman's concern, and the concern of many philosophers who discuss folk psychology, is with folk psychology construed as a *theory*, in some demanding sense of that word, and with the putatively deleterious effect that the collapse of the theory would have on the validity of our notion of propositional attitudes. By contrast, I do not claim that a commonsense model of mind constitutes a theory in such a sense, and I do not claim that the conclusions about a particular mind that one might draw on the basis of a commonsense model are anything other than defeasible suggestions that act heuristically to guide inferencing. Having said that, it would not even matter if the commonsense models *were* to be classed as theories in a demanding sense. This is because they are not meant to be literally valid, or viewed as a such by agents who think in terms of them, in the first place. I am not claiming, for instance, that when someone views someone else's mind as a CONTAINER, that the former thinks that the latter's thoughts are *really* concrete objects in different positions in his or her head. Far from being in any conflict with Goldman, my focus on commonsense models of mind actually used by people is in strong accord with his advocacy of the "psychologization" of propositional attitude research.

ACKNOWLEDGMENTS

I have benefited from discussions with Alvin Goldman, Margarita Gonzalez, Robert Gordon, Stephen Helmreich, Eric Iverson, and Yorick Wilks.

REFERENCES

Barnden, J.A. (1986). Imputations and explications: representational problems in treatments of propositional attitudes. *Cognitive Science, 10* (3), pp.319–364.

Barnden, J.A. (1989). Belief, metaphorically speaking. In *Procs. 1st Intl. Conf. on Principles of Knowledge Representation and Reasoning* (Toronto, May 1989). San Mateo, Calif.: Morgan Kaufmann. pp.21–32.

Barnden, J.A. (1990). Naive Metaphysics: a metaphor-based approach to propositional attitude representation. (Unabridged version.) *Memoranda in Computer and Cognitive Science*, No. MCCS–90–174, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Barnden, J.A. (1991). Some interplay between metaphor and propositional attitudes. In *Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts and Implicature* (24 August 1991, Sydney, Australia), pp.1–11. Proceedings printed as Technical Report CU-CS-550-91, Department of Computer Science, University of Colorado at Boulder, Colorado.

Barnden, J.A. (1992a). Connectionism, generalization and propositional attitudes: a catalogue of challenging issues. In J. Dinsmore (ed), *The Symbolic and Connectionist Paradigms: Closing the Gap.* Hillsdale, N.J.: Lawrence Erlbaum. pp.149–178.

Barnden, J.A. (1992b). Belief in metaphor: taking commonsense psychology seriously. *Computational Intelligence, 8* (3), pp.520–552.

Blank, D.S., Meeden, L.A. & Marshall, J.B. (1992). Exploring the symbolic/subsymbolic continuum: a case study of RAAM. In Dinsmore, J. (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap.* Hillsdale, N.J.: Lawrence Erlbaum. pp. 113–148.

Chalmers, D.J. (1990). Syntactic transformations on distributed representations. *Connection Science, 2* (1 & 2), pp.53–62.

Chalupsky, H. (1992). Using hypothetical reasoning as a method for belief ascription. In *Working Notes of Symp. on Propositional Knowledge Representation,* AAAI Spring Symposium Series, Stanford University, March 1992. pp.35–42.

Chellas, B.F. (1980). *Modal logic.* Cambridge University Press.

Creary, L. G. (1979). Propositional attitudes: Fregean representation and simulative reasoning. *Procs. 6th. Int. Joint Conf. on Artificial Intelligence* (Tokyo), pp.176–181. Los Altos, Calif.: Morgan Kaufmann.

Davis, E. (1990). *Representations of commonsense knowledge.* San Mateo, Calif.: Morgan Kaufmann.

Dinsmore, J. (1991). *Partitioned representations: a study in mental representation, language processing and linguistic structure.* Dordrecht: Kluwer Academic Publishers.

Fauconnier, G. (1985). *Mental spaces: aspects of meaning construction in natural language.* Cambridge, Mass.: MIT Press.

Gibbs, R.W., Jr. & O'Brien, J.E. (1990). Idioms and mental imagery: the metaphorical motivation for idiomatic meaning. *Cognition, 36* (1), pp.35–68.

Goldman, A.I. (1991). The psychology of folk psychology. Manuscript, Department of Philosophy, University of Arizona, Tucson, AZ. Presented at *17th Annual Meeting of the Society for Philosophy and Psychology,* San Francisco State University, San Francisco, Calif., June 1991.

Haas, A.R. (1986). A syntactic theory of belief and action. *Artificial Intelligence, 28,* 245–292.

Lakoff, G. (1987). *Women, fire and dangerous things.* Chicago: University of Chicago Press.

Lakoff, G., Espenson, J. & Goldberg, A. (1989). Master metaphor list. Draft manuscript, Cognitive Linguistics Group, University of California at Berkeley.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Moore, R.C. (1973). D-SCRIPT: A computational theory of descriptions. In *Advance Papers of the Third Int. Joint Conf. On Artificial Intelligence,* Stanford, Calif, pp.223–229. Also in *IEEE Transactions on Computers, C-25* (4), 1976, pp.366–373.

Reddy, M.J. (1979). The conduit metaphor — a case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought,* Cambridge, UK: Cambridge University Press.

Wellman, H.M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.