Uncertain Reasoning About Agents' Beliefs and Reasoning

John A. Barnden

School of Computer Science
University of Birmingham
Birmingham B15 2TT
United Kingdom

December 9, 2011

For special issue of *Artificial Intelligence and Law.* Revised version.

Abstract. Reasoning about mental states and processes is important in various subareas of the legal domain. A trial lawyer might need to reason and the beliefs, reasoning and other mental states and processes of members of a jury; a police officer might need to reason about the conjectured beliefs and reasoning of perpetrators; a judge may need to consider a defendant's mental states and processes for the purposes of sentencing; and so on. Further, the mental states in question may themselves be about the mental states and processes of other people. Therefore, if AI systems are to assist with reasoning tasks in law, they may need to be able to reason about mental states and processes. Such reasoning is riddled with uncertainty, and this is true in particular in the legal domain. The article discusses how various different types of uncertainty arise, and shows how they greatly complicate the task of reasoning about mental states and processes. The article concentrates on the special case of states of belief and processes of reasoning, and sketches an implemented, prototype computer program (ATT-Meta) that copes with the various types of uncertainty in reasoning about beliefs and reasoning. In particular, the article outlines the system's facilities for handling conflict between different lines of argument, especially when these lie within the reasoning of different people. The system's approach is illustrated by application to a real-life mugging example.

1 Introduction

This article is not about providing a formal model of juridical proof. Rather, it is about other types of inferential process conducted by (a) the participants in situations that come to the attention of the law, and/or (b) lawyers, judges, jurors, witnesses, police officers, and so forth. In particular, it focuses on scenarios where individuals are reasoning, uncertainly, about the mental states and processes of other individuals. Individuals' mental states are of course of direct importance in law, because for one thing laws often bring in questions of intention, sometimes under the guise of "good faith" and "malice aforethought." (See, e.g., Nissan *et al.* 1992, where a defendant's intentions are an important element in an excuse for apparently reprehensible behavior.) The article does not address all mental states and processes, but instead confines attention to beliefs and reasoning acts. Here are some example scenarios.

- A trial lawyer might wish to reason about the (conjectured) beliefs of the members of the jury about the (conjectured) beliefs of a defendant. The purpose of the lawyer's reasoning could be to see how best to try to influence the jury on some matter.
- A trial lawyer might wish to reason about a prospective juror's (conjectured) beliefs in order to decide whether to challenge the inclusion of the juror.
- A police officer might wish to reason about the (conjectured) beliefs and reasoning of an unknown perpetrator in order to come up with a good investigation strategy.
- A judge might wish to reason about the (conjectured) beliefs and reasoning of the defendant in deciding upon a sentence.
- Furthermore, in some types of case it has been reported in the media that jurors have considered what sentences the judge is likely to deliver under various different circumstances. It could be important for the lawyers in the trial to reason about jurors' reasoning about judges' reasoning about sentences.

I will use the term "agent" to mean any person—or, for that matter, AI system—whose mental states or processes are at issue. I will use the term "mental-state reasoning" to mean reasoning about the mental states and processes of agents. I will almost exclusively be concerned in this paper with the special case of "belief reasoning," by which I mean mental-state reasoning where the mental states are beliefs and the mental processes are reasoning acts leading to those beliefs. The reason for this restriction is that most of the detailed research underlying this article has been confined to belief reasoning, even though other types of mental state reasoning are of course fundamental both for general purposes and for purposes of application to legal matters.

The question for AI that is directly addressed in this paper is how to automate belief reasoning—or, less ambitiously, provide automated assistance with belief

reasoning—whether this reasoning is about the beliefs and reasoning of participants in a trial or about the beliefs and reasoning of participants in the situation a trial is about. Another question for AI is automated understanding of trial records, perhaps for the purpose of helping lawyers in further trials. Understanding of any text involving the description of people's actions (including the actions and utterances of participants in a trial) needs to "read between the lines" by coming to plausible conclusions about the beliefs and other mental states of the people involved.

The potential benefit of an AI study of belief reasoning in law is not only the development of practical working systems for law at some future point, but also the generation and moulding of new psychological investigations into how people actually reason, and additional understanding of the intricacies of reasoning about beliefs and reasoning in the context of law. There are also benefits to AI at large because of the richness of legal proceedings as an application domain.

The reasoning scenarios that are encompassed by the above comments go far beyond consideration of legal evidence under any tight or official definition of that term, because they include not only reasoning by legal non-experts but also reasoning by legal experts about matters that are not subject to rules of evidence in the first place — matters such as how jurors are thinking. However, the scenarios can also have much to do with legal evidence. Allen (this volume) places great weight on the idea that in the jury-room the real evidence is not what is produced at trial but rather the result of the jurors' deliberations on what is produced, where those deliberations are affected by the jurors' rich bodies of attitudes and beliefs. Allen points out that individual jurors' mental states can be idiosyncratic: different from juror to juror, and different from the beliefs and attitudes of the lawyers, judges, etc.[1]

Allen (this volume) also comments that "each individual juror's own knowledge is brought to bear on the questions at hand, and may in fact be discussed openly by the jury, in that sense becoming evidence in the case." Clearly, therefore, an agent outside the jury wishing to reason about conjectured jury deliberations needs to take into account jurors' reasoning about each other's beliefs; and such considerations are about legal evidence in Allen's broad conception.

Whatever the nature of juridical proof, when the decision rests on a jury the trial lawyers must try to reason about how *the jury members* reason about the information proffered in the trial, irrespective of how closely that reasoning follows the canons of juridical proof. Thus, the lawyers' handling of the evidence in the trial must bring in reasoning about ordinary people's reasoning. At the same time, lawyers may wish to consider the extent to which their own comments to the jury have affected the way the jury members reason about the evidence. A rather more special consideration is addressed by Allen (this volume) when he comments that judges in the U.S.A. can exclude relevant evidence if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury (in Federal Rule of Evidence 403). But these dangers are to do with the belief states and reasoning of the people who might be prejudiced, etc. In addition, a lawyer may need to reason about how a judge

might reason about such prejudice, etc.

A major technical concern of the present article is the uncertainty inherent in reasoning about beliefs and reasoning. Indeed, the legal domain provides good illustrations of why AI accounts of mental-state reasoning need to incorporate a powerful treatment of uncertainty. Unfortunately, the research in various disciplines on mental-state reasoning has largely skirted the question of uncertainty of reasoning, at least when it comes down to detailed technical proposals. But in practical applications — notably including law — conclusions generally do not follow with complete certainty from premises. This is, of course, well-known, and is emphasized by Allen (this volume). In short, straightforward use of standard deductive reasoning is of limited interest, and one must allow for levels of uncertainty in hypotheses, and, relatedly, for retractability (defeasibility) of tentatively established hypotheses when evidence mounts against them. Research into uncertain reasoning within AI has been conducted largely separately from research into reasoning about agents' beliefs (though see Konolige 1985, 1986, 1988, Ballim & Wilks 1991, Perrault 1990, Cravo & Martins 1993, Asher & Lascarides 1994, Dragoni & Puliti 1994, Kaplan & Schubert 1997a,b; see also Parsons, Sierra & Jennings 1998 for related work).[2]

The introduction of uncertainty into mental-state reasoning greatly complicates the latter. Part of the main purpose of the present article is to present some of the complications, and to briefly report the approach I have been taking to them in an implemented system for belief reasoning, called ATT-Meta (Barnden *et al.* 1994, 1996, Barnden 1998a,b).[3] To my knowledge no other system addresses these complications.

It is as well to say what this article does not purport to do. It does not purport to present a reasoning system that is adequate to all the subtlety and richness of the reasoning scenarios presented above. The ATT-Meta system is but a step towards such adequacy, and it is the principles and computational issues informing the development of the system that are of prime interest, not the system as such. There is no attempt in the system to consider legal rules about how to take the mental states of defendants (etc.) into account. The system's handling of uncertainty is fairly primitive, even though as an uncertain reasoner *about beliefs and reasoning* it is relatively advanced. In particular, it does not fully cope with a point made by Allen (p.c.) to the effect that a mass of discrete pieces of evidence each of which is not very persuasive in itself can add up to a very persuasive overall story. Finally, Allen (p.c.) states that much of the process in trials is to do with establishing burdens of proof rather than matters of fact. The present article does not address the issue of burdens of proof, and in any case, as stated above, the article is not in any case directly addressed at the question of how to prove things in court.

The rest of the article is structured as follows. Section 2 describes a real-life mugging example that will be used at various points later. Sections 3 and 4 discuss the rich ways in which uncertainty enters into belief reasoning, including when applied to the law-based scenarios of the sorts alluded to at the start of this Introduction. Section 5 discusses one of the major computational complications arising

from including uncertainty-handling in belief reasoning, *viz* the danger of a severe combinatorial explosion of reasoning subgoals. Section 6 makes a case for qualitative uncertainty methods as opposed to numerical statistical ones. Sections 7 and 8 describe one of the main techniques in ATT-Meta for belief reasoning, *viz* simulative reasoning, which is well-known, but which is considerably elaborated in ATT-Meta because of thorough inclusion of uncertainty-handling. Section 9 notes some basic technical points about ATT-Meta needed for the remaining sections. (The article is not intended as a complete description of ATT-Meta. Additional information can be found in Barnden *et al.* 1994, 1996 and Barnden 1998a,b.) Sections 10 and 11 address a second major complication introduced by including uncertainty in belief reasoning, *viz* the need for properly coordinating conflict resolution (between competing lines of reasoning) across different layers of belief. Section 12 describes in detail how ATT-Meta applies the methods discussed in previous sections to the mugging example in Section 2. Section 13 concludes.

2 A Real-Life Example

Here I present an example that gives, in a small way, an idea of the intricacies involved in reasoning about beliefs and other mental states in situations of interest to Law. We will use the example in later sections to illustrate various general points, as well as to see in a specific case how the ATT-Meta system reasons.

The example is based on a real-life but fortunately minor mugging event in which I was the victim. Using this event in this paper will enable us to consider both the actual inner perspective of someone involved in a criminal scenario as well as the perspective of people thinking about the scenario from the outside. While I was walking on a street next to a railway station in a big city at about 8pm, two youths jumped upon me from behind. I struggled while they held my arms and ripped my wallet from my back pocket, destroying my trousers in the process. In the course of the struggle I fell on one knee to the ground, bruising it. The interesting thing about the event was that I had the definite impression that the youths were trying *not* to hurt me. I cannot, however, give a clear reason for this impression. The youths certainly did not use or threaten me with violence (other than by virtue of grabbing my arms etc.), or appear to have any sort of weapon. That I thought the youths were trying not to hurt me became a part of my police statement.

For the purposes of the paper I will simplify in various ways. I will reduce the number of perpetrators to one, and will use the artificial name "Perp" for him. I will call the victim "Vic" rather than referring to myself. I will consider the proposition that "Perp did not try to hurt Vic during the mugging" as opposed to the stronger proposition that "Perp tried not to hurt Vic during the mugging."

One point of mentioning the example is to raise the question of how a fact-finder (e.g. police officer) who is considering the event could come to the plausible conclusion that Vic did indeed *believe* that Perp did not try to hurt him during

the mugging. This would have to be inferred somehow from the fact that Vic *said* that he believed that Perp did not try to hurt him during the mugging. One complication is (let us assume) that the fact-finder him/herself believes that muggers normally try to hurt their victims, and attributes this belief also to victims; so that without Vic's statement about lack of an attempt to hurt, the fact-finder would have assumed that Vic believed that Perp was trying to hurt him. Of course, the idea that muggers normally try to hurt their victims is an over-simplification (of reality, and as a portrayal of my (Vic's) own beliefs), but we will stick to it for the sake of illustration. Another complication is that, naturally, a fact-finder should not unreflectingly believe what a victim states. For all the fact-finder knows, Vic is lying in order to protect Perp for some reason.

From the above conclusion that *Vic believed* that Perp did not try to hurt him during the mugging, the fact-finder might make the further plausible inference that Perp did not *in fact* try to hurt Vic, as an exception to the default that muggers do try to hurt their victims. We will look at how a fact-finder could make that inference.

The question of whether Perp in fact tried to hurt Vic could be important in considering whether Perp is responsible for the (minor) injury to Vic's knee. However, as this brings in detailed matters of actual law about robbery, etc., with which I am unfamiliar, I will not consider it in depth. The reader should keep in mind that this example is not presented as a contribution to the understanding of how to reason about *mugging* or other particular crimes or misdemeanours, but rather to the understanding of reasoning about *people's beliefs and reasoning about* events of legal interest.

Now, the fact-finder could be trial lawyer, judge, juror, insurance agent, etc. rather than a police officer. (The actual mugging did not feature in a trial or insurance deliberations, but we can extend the example in our imagination.) We then get, as well as the question of a lawyer (or whoever) reasoning about the mugging situation and the beliefs therein, the question of, say, a lawyer trying to reason about, say, *jurors' reasoning* about the mugging situation and the beliefs therein. That is, the whole reasoning-by-fact-finder scenario we were previously considering can be "embedded" as something reasoned about within the mind of another person (e.g. a juror). It could be important for a lawyer to reason about the jurors' reasoning in order to determine how best to persuade the jury.

Of course, an additional complication here is that jurors' may have particular beliefs that are different from beliefs of other jurors or from beliefs of the lawyer. (Allen, this volume, gives an important role to this point.) The lawyer may or may not know about some of these beliefs. In the absence of specific information about relevant beliefs, the lawyer has to make default assumptions about them (e.g., the lawyer might assume that the jurors believe that muggers normally try to hurt their victims).

So far the inferences mentioned have all been those of imagined *human* agents such as police officers, lawyers, jurors, etc. The AI question is how to develop an automated system that could make some of the inferences discussed. For example, the system might reason about jurors' reasoning about Vic's beliefs, as an aid to

a human lawyer. Notice in particular that, even though the automation of jurors' reasoning is not directly a goal, automation of reasoning *about* such reasoning is a legitimate goal.

3 Belief Reasoning and Uncertainty

Many forms of uncertain reasoning have been studied in AI. Textbooks such as Davis (1990), Rich & Knight (1991) and Russell & Norvig (1995) provide a good overview between them. Included under uncertain reasoning are default reasoning (classic example: birds fly by default; penguins don't fly; Bobby is a bird; Peter is a penguin; so Bobby presumably flies, but Peter doesn't fly), abduction (basically, inferring reasons, causes or explanations from symptoms or effects), induction, analogy-based reasoning (especially in the form of case-based reasoning), and various forms of probability-based methods, notably methods based on Bayes' rule. All these types of reasoning are relevant to the present paper, either in that they enter into court participants' reasoning or they enter into the reasoning of participants in situations being adjudicated. Case-based reasoning has been widely emphasized in AI and Law, and uncertainty more generally is widely recognized as an important concern. For instance, Allen (this volume) places much weight on various forms of uncertainty in law.

The purpose of the present section is, however, more specific, *viz* to point out the variety of ways in which *belief reasoning* is inherently uncertain, whether applied in the legal area or elsewhere.

First, it is obvious that people do not actually draw all possible conclusions from their existing beliefs. Therefore, normally, an agent X should not conclude that it is *definitely* the case that an agent Y believes R simply because (according to X) Y believes some propositions P_i, where R follows from the P_i. Thus, reasoning about beliefs should immediately take one into the arena of uncertain reasoning. Conclusions such as that Y believes that R must usually in practice be marked as uncertain, and must usually be amenable to being retracted—or, more generally, to having their level of certainty decreased—because when a conclusion like Y-believes-R is initially formed the evidence against it may not yet have been uncovered. (Here and henceforth I use the term "evidence" mostly in its everyday meaning, not in any restricted legal sense.)

There are additional reasons why belief reasoning is uncertain. Even if X were absolutely certain that Y has pursued some argument that culminates in a conclusion R, X should not be certain that Y believes R, because for all X knows Y may also have pursued some other argument that supports NOT(R), where the latter argument is the stronger in Y's view. The argument may not be one that X is equipped to realize that Y is capable of pursuing, either because of the nature of the reasoning steps, or because X is unaware that Y believes some of the propositions used in the argument.

A different consideration is that, whether or not Y has pursued an argument that culminates in a proposition R, there may be evidence that it is not the case that

Y believes R. For instance, a reliable informant (perhaps Y him/herself) may have told us that Y lacks the belief R. More generally, some chain of reasoning or other might provide evidence that it is not the case that Y believes R.

When Y is supposedly basing inferences on premises P_i, there is the question of how X comes to ascribe those premises to Y in the first place. The evidence supporting this ascription may involve uncertain reasoning procedures (such as default reasoning or reasoning by analogy), or may be based on uncertain information. The evidence that Y believes a P_i might be that Y is a certain type of person (e.g., a Labour Party supporter), but in such a case it could be merely a default that that type of person believes P_i, and the information that Y is indeed of that type could be uncertain. (E.g., the proposition that Y is a Labour Party supporter might be based on a document whose authenticity is doubtful.) Alternatively, the evidence that Y believes some P_i might be that some informant, W, has said so. But X might choose to retain some doubt that W is correct (cf. the comments in Schum (this volume) on evidence about a trial witness's veracity, objectivity, observational sensitivity and competence). Or, W's statement might be expressly uncertain, as in "Y seems to believes that R." Another type of support for the hypothesis that Y believes P_i is that Y has stated P_i. Depending on circumstances, and knowledge of Y's veracity, etc., one may plausibly infer that Y believes P_i.

There is also a rather different connection between uncertainty and belief reasoning. Most discussions of it, at least within AI, focus on the case where R follows from Y's other beliefs P_i by classical deduction (e.g., modus ponens steps, or resolution steps). But, clearly, if the reasoning agent X itself[4] is capable of types of inference other than classical deduction, when reasoning about things in general, then X ought to be able to view other agents Y as doing those other types of inference. For instance, if X is capable of doing induction, it should surely be able to cast other agents as doing induction. The same applies to other forms of uncertain reasoning such as abduction, default reasoning and analogy-based reasoning.[5]

Thus, in the scenario about Y reasoning to R from some propositions P_i, Y's own alleged reasoning may be uncertain. Y may be using induction, abduction, default reasoning or analogy-based reasoning. Indeed, since completely-certain reasoning is of relatively little interest for real applications (since little is certain in the real world), *most* of the alleged reasoning of Y will, in practice, be uncertain. Notice that, as a special case of this, Y itself may be reasoning about the beliefs of a further agent Z, and that the results of this reasoning are therefore uncertain for all the reasons so far discussed. As another special case, Y may be reasoning about what the law commands on a particular topic. But notice that, as Allen (this volume) says, there are "endless exceptions" in the "web of [legal] regulation."

Altogether, then, when X is reasoning about Y's reasoning, there are two layers of uncertainty to worry about: the uncertainty inherent in the reasoning steps attributed to Y; and the uncertainty inherent in the question of whether Y has actually performed available inference steps (whether these are uncertain or

not), and whether Y believes the conclusion of the steps, even given that Y has performed them. These two layers of uncertainty are largely independent—for instance, X could be very confident that Y has done some particular inference steps that are laden with uncertainty, or be very uncertain as to whether Y has done some particular high-certainty inference steps.

It is worth noting that to reason that an agent Y reasons from some propositions P_i to some conclusion R is but one way in which one might conclude that Y believes R from the premise that Y believes the P_i. Let us call that premise YP, and use YR to mean the proposition that Y believes R. There could be arguments that establish YR on the basis of YP that do not rely on considering Y's own reasoning. For instance, the reasoner might just be aware of a strong, observed correlation between believing the P_i and believing R. A system for evaluation of jury members might contain the rule:

IF person p believes that strikers should be put in prison

THEN (to some degree of certainty) p believes that welfare beneficiaries are swindling the country.

One does not need to know of any reasoning (by person p) underlying such a correlation in order to be able to exploit the correlation. Also, the argument from YP to YR might go through propositions that are not about Y's mental states at all. From the hypothesis YP that Vic believes his Platinum Mastercard was stolen during the mugging one might plausibly conclude that his Platinum Mastercard was in fact stolen during the mugging. From this one can, naturally enough, conclude that he had a Platinum Mastercard account. This could then lead to a conclusion YR that he believes that it is beneficial to have such an account.

The ways in which one might conclude that agent Y believes something R that do not involve considering Y's own (alleged) reasoning are important, because they show that belief reasoning is inextricably entwined with ordinary reasoning about the world, and that there can be radically different types of evidence for propositions about people's beliefs.

4 Reasoning Omissions

Reasoning omissions are an important topic for law and related areas such as ethics. By a reasoning omission I mean a case where agent Y does not draw conclusion R from propositions P_i that Y (supposedly) believes, where R follows from the P_i in some way, at least in the view of an agent that is reasoning about Y's reasoning. Thus, omission is a view-relative matter.

Also, in saying that R follows from the P_i we are not assuming that the following is by any sort of definite (i.e., certain) reasoning scheme, such as classical deduction. R might (allegedly) follow from the P_i by abduction or by analogy-based reasoning, say. Because of this and the view-relativity, it may be that R is in fact false even though the P_i are true, so that Y would, whether by luck or by superior

knowledge or reasoning ability, be correct in not concluding R. However, we will still say that Y has "omitted" some reasoning.

An agent Y can "omit" to apply reasoning steps for any number of reasons. The subject matter may be unfamiliar, and Y may therefore not apply a reasoning method she is perfectly capable of using for other subject matters. Y may be exhausted, sleepy, drunk, drugged, nervous, sexually aroused, insane, distracted by other concerns, stupid, hurried, or endowed with superior knowledge about the situation. Y may not, metaphorically speaking, have "brought the beliefs P_i together in her mind" — a very common phenomenon, especially when different P_i are about different domains of life. See Barnden *et al.* (1994, 1996) and Barnden (1998a,b) for discussion of metaphors of mind and for the ATT-Meta system's reasoning from metaphorical statements about mental states.

Clearly, if Y fails to observe that something follows from something, that may excuse her from being thought to have deliberately done something wrong. For instance, if she fails to see that using a certain word she has spoken is liable to cause offence to some person Z, perhaps because she has not made some inferences about what race Z belongs to, or failed to infer that Z could hear what she was saying, then she cannot reasonably be accused of deliberately offending the person. Of course, this is an ethical conclusion: the law may choose to ignore the question of deliberateness. When deliberateness is relevant, though, a legal or ethical judgment might take into account the reason for the reasoning omission. If the reason is that Y is drunk, then that may be in itself reprehensible enough — enough under Y's control — for Y still to be considered reprehensibly responsible for the offence to Z. If the reason is that Y was distracted by a death in the family, then perhaps Y would not be considered reprehensibly responsible for the offence to Z. These concerns touch upon a large literature on whether people can be said to intend the expectable side-effects of the actions they intend — see, e.g., Bratman (1992).

For the purposes of the present paper, I wish merely to note that a practical belief-reasoning system must have the ability to reason (uncertainly) for and against the occurrence of reasoning omissions, and must be able to ascribe such reasoning about reasoning omissions to other agents, where the nature of omission itself depends on those agents' beliefs.

As an added complication, reasoning about reasoning omissions is itself a form of reasoning that could be subject to omission. The latter omission could be called a meta-omission. Lest all this should seem impossibly abstruse, consider the following situation. Wife W of pathologically jealous husband H talks to handsome man M at a party. This leads to making H batter W, because he assumes that she should reason that talking to M would upset H, and therefore thinks that W talked to M in full knowledge that it would upset H. The reasoning that H ascribes to W has not, however, been done by W in actuality. Thus, W has "omitted" to perform a reasoning act that H assumes she has performed. H has omitted to reason that W "omitted" that act, which is one that neither we nor the wife would have viewed it as a reasonable act in the first place. So, we have a view-relative meta-omission.

5 Belief Nesting and a Bomb Threat

For the present section, two detailed points about belief are crucial.

First, is simplistic to talk about X taking Y to believe something R: rather, X has some level of certainty that Y believes R to some level of certainty. The two layers at which certainty can vary are independent. Obviously, with more layers of belief one has more layers of certainty variation. (Nevertheless, I will often talk about belief without mentioning levels of certainty, in the interests of brevity.)

Second, it is important to bear in mind the distinction between NOT(Y-believes-R) and Y-believes-NOT(R). In English, we could respectively say, "Y lacks the belief that R" and "Y believes that it's not the case that R." Notice carefully that Y may lack a belief both in R and in NOT(R). That is, both NOT(Y-believes-R) and NOT(Y-believes-NOT(R)) could have high or perfect certainty. In this case Y simply doesn't have a view on the matter of R. The distinction in question is obscured in English and other languages by the linguistic phenomenon of *raising*, whereby, say, "Y believes that Tom won't be coming" is often expressed as "Y doesn't believe that Tom'll come" even though a pedantic reading of the latter would make it mean merely that Y lacks the belief that Tom is coming, saying nothing about whether Y believes that Tom is not coming.

Now, in doing uncertain reasoning, whether about beliefs or not, the investigation of a hypothesis Q often involves considering NOT(Q) as well, to take account of the possibility that the evidence for NOT(Q) is at least as strong as the evidence for Q. Now consider a Q that may be a belief of an agent Y. By what has just been said, there is also NOT(Q) to consider "within Y." That is, Y's own reasoning towards both Q and NOT(Q) must (often) be considered. But, the hypotheses Y-believes-Q and Y-believes-NOT(Q) may be supported by arguments outside Y (i.e., arguments that do not involve considering Y's own reasoning). But therefore we may have NOT(Y-believes-Q) and NOT(Y-believes-NOT(Q)) to consider outside Y as well. In sum, for each proposition inside Y we may have *two* propositions outside.

But this does not take levels of certainty at different layers into account. Let us assume that there are finitely many levels of positive certainty that a proposition can have. (This accords with ATT-Meta's method, but in any case the complications to be described would be worse if certainty values lay on a continuous range.) Let n be the number of levels. Then instead of just Y-believes-Q we actually have n different propositions Y-believes-Q-with-certainty-δ, for the different values of δ. Then there are the negations of all these propositions. So, altogether, we have $2n$ propositions outside Y for each proposition Q inside Y.

And this applies to every layer of belief. So, if we are considering whether Z believes that Y believes that Q, there are $4n^2$ propositions to consider outside Z. Here we see a serious combinatorial explosion. In ATT-Meta, n is 4, so the multiplication factor per layer of belief is $2n = 8$. A major task within the development of ATT-Meta has been to devise optimizations to economize on this number. Fortunately, the optimizations lead to an factor as low as 1.17

in favourable (yet still non-trivial) cases, and rising to usually no more than about 1.7 in cases of typical complexity. (These numbers are averages, within a single system run, over all hypotheses that are within some agent's belief set.) The highest value I have seen in experiments with the most recent version of ATT-Meta is 2.0, except for an anomalous experiment in which the believing agent had contradictory, completely-certain beliefs, and in which the explosion ratio was 6.0 (though only across one boundary between belief layers). Even 2.0 is very good compared to what would happen without the optimizations. For reasons of length we will not consider the optimizations adopted. However, they are outlined in Barnden (1997).[6]

The multiplication factor just discussed does not take account of all hypotheses. In particular, it ignores the special "agent-inference" hypotheses introduced below in section 11.3. If these are considered and the explosion factor across a belief boundary is calculated as the ratio of the number of hypotheses on either side, the resulting explosion factor runs from 1.29 in favourable non-trivial cases to about 1.8 in typically complex cases, with an observed maximum of 2.63 (except for the anomalous case mentioned above, where the factor was 4.0 across the single boundary). However, it should be emphasized that in ordinary circumstances the agent-inference hypotheses lead to very little inference activity, so that the statistics in the previous paragraph may give a fairer picture of the explosion in practice.

It is important to notice the role of uncertainty in generating the combinatorial explosion. If all reasoning were certain, then if one established that Y's own (alleged) reasoning supported Q there would be no need to look at Y's reasoning towards NOT(Q) (unless one wanted to take account of directly contradictory, certain beliefs); hence there would be no need to look at Y-believes-NOT(Q) or its negation. Also there would be no need to look at NOT(Y-believes-Q), because Y-believes-Q would be certain. And without uncertainty there would be no need for different hypotheses of the form Y-believes-Q-to-degree-δ for different values of δ.

6 Quality not Quantity

Russell and Norvig (1995) rightly point out that probabilistic reasoning can often deliver a worthwhile result where qualitative reasoning methods such as default logic are in danger of yielding no result or a misleading result. However, although those authors are aware that the necessary numbers (i.e., prior and conditional probabilities) on which to base the reasoning may be difficult to ascertain, they fail to note the yet more serious problem that for some broad applications of uncertain reasoning the numbers don't even *begin* to be available—one has no hope whatsoever, in practice, of obtaining them, or at least obtaining them in time to do the desired reasoning.

The main example comes from the uncertain reasoning that one might do with entirely qualitative information conveyed by natural language input; and this has

major implications for the legal domain. Suppose, for instance, that S says to H: "Usually it is illegal to kill bats in your home." Then, surely, we expect H to be able to infer that (presumably) his action of killing the bat hanging at the end of his bed was illegal. We do not expect H to insist on being given any sort of numerical measure of such an act being illegal before he is prepared to form the conclusion. In short, a person H is capable of dealing with entirely qualitative uncertain information conveyed in natural language sentences, even if the subject matter is unfamiliar (so that H has no basis on which to summon up statistics). Therefore, it is desirable for AI systems to be able to do the same thing, and, importantly for the present paper, *to be able to reason about people doing it.*

The bat example rested on a consideration of natural language input. Reasoning based on such input is important for some types of AI application to law. For instance, an AI system might be required to reason about what inferences people, including jurors, witnesses and defendants, can be expected to draw from natural language sentences they hear. Also, an AI system that examined legal statutes or court cases should be able to reason with default principles involved in them (such as a principle that a certain type of evidence is normally inadmissible), without having to rely to any numerical measures of their uncertainty.

Furthermore, if a system is to reason about the reasoning of some other agent X (juror, lawyer, defendant, witness, or whomever), it should exploit information about *X's view* of the world, gleaned from X's natural language statements. That view may contain uncertainties that the system is simply in no position to adumbrate numerically. Suppose witness X states a default such as that "John usually goes to London on weekends." On the hypothesis that X is not lying, inferences about X's beliefs about John's whereabouts on a given weekend should take the default into account. Now, it may well be impossible to find out from X the proportion of weekends that John is actually in London. X may not know the proportion, and the default may be just a vague impression X has. Another possibility is that X has heard the default from someone else who, in turn, failed to communicate any numerical measure of the default. And it goes without saying that any observation of John's actual behavior would be irrelevant, as what is at issue is X's beliefs about John, not the reality about John.

Quite apart from an agent X's own uncertainty in believing a proposition, there is the the uncertainty another agent Y will generally have about whether X has any specific belief (to whatever level of certainty). In practice it is extremely unlikely, at least for some important types of beliefs, that Y would know or be able to wok out he probability that X has the belief. The world is too complex to imagine that one would ever have enough evidence or experience to make such a determination, especially in the relatively unusual circumstances that are addressed in legal proceedings. The problem is even worse in the case of nested belief. On the unusualness of circumstances, compare some comments Schum (this volume) makes, such as "[in law] we usually encounter singular, unique, or one-of-a-kind events for which no meaningful statistical analyses are possible."

Russell and Norvig (1995) suggest at one point (p.458) that the human brain may use quantitative uncertainty handling for propositions (at, say, the neural

network level). However, this is only a speculation, and in any case, for AI purposes, it remains to be demonstrated that cashing out inherently qualitative uncertainty measures (such as the English word "usually" in the bat example) as necessarily-arbitrary numerical measures is any better than using qualitative measures directly.

Henceforth I will only explicitly consider qualitative uncertainty, especially as this is the only type included in the ATT-Meta system. However, I am certainly not claiming that qualitative uncertainty is the only type that should ever be used, and many of the points to be made can presumably be adapted to encompass quantitative uncertainty as well. Also, this article makes occasional mention of analogy-based reasoning (and case-based reasoning), which are often and crucially enriched with numerical measures. For instance, numbers are often used for measuring the strength of an analogy or case-match. Finally, as Schum (this volume) points out at the end of his article, qualitative models can be useful in guiding later probablistic analyses when these do become possible.

7 Simulative Reasoning

Suppose that, according to agent X, agent Y believes some propositions P_1 to P_n. Suppose further that X thinks that R can be inferred from the P_i. Then, it is (very often) reasonable for X to at least tentatively suppose that Y has inferred R, and therefore for X to at least tentatively suppose that Y believes it. The question is, how exactly is X to come to this conclusion about Y?

The paper focuses on *simulative* reasoning as the main answer to this. Intuitively, when an agent X engages in simulative reasoning about an agent Y, X "stands in Y's shoes" and tries to reason as if she, he or it were Y, using beliefs that X believes Y to have. Another way to put it is that X pretends to adopt beliefs that X conjectures Y to have, and reasons on the basis of them rather than on the basis of X's own beliefs. During the reasoning, X uses its own reasoning rules, strategies or whatever.

To make this more definite, let's look at the case where n is 2 and P_2 is $P_1 \Rightarrow R$ (where $\Rightarrow$ means material implication), so that R follows from the P_1 and P_2 by a modus ponens step. Let's also suppose that X represents Y's belief in the P_i by the formulae

`bel(Y,` P_1`)`

`bel(Y,` $P_1 \Rightarrow R$`)`

with P_i and R replaced by the particular formulae that they are.

The simulative reasoning approach is that X's reasoning system constructs a computational environment in which P_1 and P_2 play the role of premises. This action partly consists in stripping off the `bel` layer from the above two formulae. Then, X infers R from P_1 and P_2 within the special environment by a straightforward application of modus ponens. The inferential act here is essentially the same as

if X were inferring R from the P_i for its own purposes. X can now re-introduce a `bel` layer round R to come up with the conclusion:

`bel(Y,` R`)`.

Clearly, the account can be generalized to cover cases in which there are more than two base propositions P_i, inference step types other than modus ponens are involved, and R follows from the P_i by a chain of several steps.

Simulative reasoning is ATT-Meta's main technique for reasoning about agents' reasoning. However, ATT-Meta can also do non-simulative reasoning about agents' beliefs. For instance, it can do the types of reasoning discussed in section 3 which do not involve consideration of Y's own reasoning.

Simulative reasoning has been a popular technique in AI for reasoning about beliefs (see, e.g., Moore 1973, Creary 1979, Konolige 1985 and 1986— where, however, simulative reasoning is atypically called "attachment," and "simulation" means something entirely different — and Haas 1986, Ballim & Wilks 1991, Dinsmore 1991, Chalupsky 1993 and 1996, Attardi & Simi 1994, Kaplan & Schubert 1997a,b). It has been popular especially with investigators interested in producing practical, working systems as opposed to theoretical frameworks. It has also been advocated by a number of philosophers of mind and cognitive psychologists (e.g., Gordon 1986, Goldman 1992, Harris 1992), though there has been intense debate on its merits (Davies & Stone 1995, Carruthers & Smith 1996).

An efficiency advantage of simulative reasoning with respect to competing techniques is discussed by Haas (1986). Barnden (1995) reviews that advantage and presents additional ones. The most important one for the present article is that it relatively easily allows a reasoner X to impute to Y *any* type of reasoning that X itself does, no matter how complex it is. For instance, it is much more straightforward for X to impute analogy-based reasoning, abduction and other forms of uncertain reasoning to Y than it is in competing approaches.

In law, simulative reasoning is often needed for simulating how a reasonable person would think, rather than simulating an actual person (Allen, p.c.). Both applications are important in the reasoning scenarios envisaged in the present article, and simulative reasoning is just as capable of simulating a hypothetical reasonable person as an actual person.

I have presented simulative reasoning as proceeding forwards from premises P_i to a conclusion R. But a backwards (i.e., goal-directed) form of simulative reasoning is also possible. If X has the proposition Y-believes-R as a reasoning goal, Y strips off the belief layer to get plain R as a goal within the simulation. Working backwards from R, X may now find subgoals P_1 and $P_1 \Rightarrow R$, say. X may now notice that it already knows that Y believes these two propositions, or may need to do further backwards reasoning to establish that Y believes them. ATT-Meta does a form of backwards simulative reasoning.

Because most discussions of simulative reasoning fail to consider uncertainty, they fail to consider the following issue. X may be very sure that Y believes some proposition P but very unsure whether Y believes Q. Should X include Q in a

simulation of Y's reasoning? This question is especially pressing where Q might support a proposition that conflicts with what P supports, particularly when Y's (alleged) confidence in Q is greater than Y's (alleged) confidence in P. (Recall that those confidence levels are entirely independent of the levels of confidence to which X holds that Y has P and Q as beliefs.) Some qualitative decision has to be made about whether to involve Q in the simulation of Y, because otherwise that reasoning would have to take care of a vast web of possible alternatives as to how Y is reasoning.

The approach taken in the ATT-Meta system is that a proposition Q that Y may believe is only included in a consideration of Y's reasoning if the hypothesis "Y believes Q" is at least a working assumption, and not just something that *might* be true.[7] This thresholding applies at every layer of belief. If ATT-Meta is reasoning directly about Y (so ATT-Meta is X in the previous paragraph), then ATT-Meta itself must take "Y believes Q" to be (at least) a working assumption for Q to be included. If ATT-Meta is reasoning about another agent X reasoning about Y, then ATT-Meta must take it at least as a working assumption that X takes it at least as a working assumption that Y believes Q. And so on in more deeply nested situations. (However, the certainty level that the innermost agent, Y, attaches to Q is not constrained.) This scheme is quite possibly too simple-minded, but at least the sheer fact that the system attempts to thoroughly combine uncertain reasoning with simulation at all is an advance.

Simulative reasoning by agent X about agent Y rests on X's use of its *own* reasoning schemes within the simulation. It therefore rests on X implicitly assuming that Y uses the same reasoning methods as X. But what if X believes that Y uses some reasoning scheme that is not in X's own arsenal, perhaps because the scheme is, in fact, faulty in X's view? This is an important consideration in that people do reason in faulty ways. Of course, there are difficult issues here concerning how X could possibly work out what reasoning schemes Y uses, short of subjecting Y to extended psychological experiments. But, in the special case where X already knows some faulty reasoning scheme S to be one that people commonly use, we can imagine X assuming, on the basis of a manageable amount of evidence of Y's behavior, that Y uses S. For instance, X might observe that Y has often formed generalizations about a whole class of people (e.g., French women) on the basis of observation of just one member of the class. X might therefore surmise, by a reasonable induction, that Y engages in unreasonable induction.

Simulative reasoning in the pure form described above naturally provides no help in such cases, because it involves X's ascribing its own reasoning schemes to Y. However, an obvious variant of simulative reasoning can still be proposed. We just allow X to use, within the simulation of Y, a reasoning scheme S that X itself does not use when not doing a simulation. Notice, however, that this does require X to formulate that scheme as a runnable procedure.

8 Mixing Simulative and Non-Simulative Reasoning

Because not all belief reasoning involves reasoning about agent's own reasoning, a practical belief-reasoning system needs non-simulative belief reasoning as well. This leads to a somewhat complex system *even if* no uncertainty is involved. In attempting to show that an agent Y believes R, the reasoning agent X needs to be able to try both simulative reasoning and non-simulative reasoning. Of course, if X tries non-simulative reasoning first and thereby proves (i.e., proves for certain) that Y believes R, there is no need for X to try simulative reasoning as well; only if the non-simulative reasoning fails does X need to try simulative reasoning. The same applies of the methods are applied in the other order.

We now come on to some of the major added complications that uncertainty of reasoning generates. Suppose X applies non-simulative reasoning to determine whether Y believes R, and computes some high but non-perfect degree of certainty for that proposition. Then it is still important for X to apply simulative reasoning. This is because, even though the results of simulative reasoning are themselves inherently uncertain, they could lend additional support for the proposition that Y believes R and therefore raise its level of certainty. In addition, evidence that arrives later or is considered later on may undermine the line of argument used by the non-simulative reasoning, but leave the simulative reasoning intact; or vice versa.

Things are yet more complicated than this, however. Suppose X thinks that R follows (uncertainly) from Q, that this follows (uncertainly) from P, and that Y believes P. X cannot straightforwardly just conclude (uncertainty) that Y believes R. Rather, X should entertain the possibility that there is evidence that undermines the hypothesis that Y believes the intermediate proposition Q. Now, the relatively simple case is that Q is undermined *within* the simulation, say because NOT(Q) follows with certainty from some S that Y believes. All we have here is more simulation. But it might equally be the case that this within-simulation undermining does not happen, but that instead X knows that Y lacks the belief Q, or X can reason non-simulatively that Y lacks it. Thus, in simulatively doing a chain of reasoning on behalf of Y, X must look at *intermediate* within-Y propositions like Q, put the Y-believes layer back, and investigate NOT(Y-believes-Q) non-simulatively. Of course, this generally entails looking non-simulatively at Y-believes-Q as well. Naturally, the situation described can arise when the X layer is itself nested as a simulation within another layer.

9 Basic Nature of ATT-Meta

This section mentions a few additional points about ATT-Meta that need to be understood for the purposes of the next section.

Propositions entertained by ATT-Meta in te cousre of reasoning are either facts or reasoning (sub)goals. They are collectively called "hypotheses." The certainty

levels that hypotheses can have are: `certainly-not, possible, suggested, presumed` and `certain,` in increasing order of strength. `Certainly-not` means that the negation of the hypothesis is certain. `Possible` merely means that the negation of the hypothesis is not certain. It does *not* mean that there is necessarily any evidence at all for the hypothesis. So, every hypothesis is possible until proven otherwise. When a hypothesis is tagged with certainty level `presumed`, it means that ATT-Meta takes the hypothesis as a working assumption (or "default"): that is, ATT-Meta will proceed for the moment as if the hypothesis is the case, but will be prepared for evidence to arise that defeats it. (Note that the concept of presumption being appealed to here is the lay, common-sense one, not the technical legal one.) `Suggested` means that the hypotheses has some evidence in its favour, but not enough for ATT-Meta to take it as working hypothesis.

ATT-Meta's set of certainty levels is simple, but surprisingly powerful in practice. The set is not meant to be the final word on qualitative certainty levels, and I envisage enriching them in the future.

ATT-Meta is a rule-based system. An ATT-Meta rule is intuitively of form

IF A_1 AND A_2 ... AND A_n THEN (with certainty C_0) B.

C_0 is always at least `suggested.` A rule is applied backwards or goal directed way: for a particular hypothesis H that is under investigation, and for a particular rule R, the system tries to match the "B" part of R to H, thus possibly instantiating variables in B. Any such variable bindings are handed back to the IF part of the rule, so that instantiated versions of the A_i are set up as subgoals. (The A_i can introduce further variables, but we will ignore that issue in this paper.)

If the A_i subgoals are satisfied to certainty levels C_i that are all at least `suggested`, then the rule contributes a certainty level C to the goal H, where C is the minimum of C_0, C_1, ..., C_n (so that the rule's own qualifier C_0 serves as an upper bound on the certainty that the rule can contribute to H.) Otherwise, the rule makes no contribution at all to H. The maximum of the certainty levels contributed by different rules to H is used as the overall rule-based contribution to H.

In example rules used below, C_0 will always be `presumed` (expressed by the word "PRESUMABLY" in the English paraphrases we will give for the rules). Each of the rules can therefore be thought of as a default rule, because no conclusion of the rule can be rated higher than `presumed` by virtue of this rule alone.

For each rule application that supports H, a *rule-application record* is attached to H. This record includes the name of the rule and the subgoals arising from the conditions A_i. When there is a need for conflict resolution (see below) between H and NOT(H), the system traces back through rule-application records, attempting to see whether one of the two hypotheses has more specific evidence than the other. Rule-application records have potential significance beyond ATT-Meta-style conflict-resolution. Schum (this volume) touches on the importance of the discovery of arguments linking evidence to hypotheses, especially within his discussion of "intellectual audit trails."

ATT-Meta has no fixed set of facts and rules, apart from som special rules for belief management and for handling logical combinations of hypotheses. Instead, the user gives the system a set of rules and facts at the beginning of a run. A current major simplification is that all content-specific rules the system uses in its own reasoning are also ascribed to all agents. That is, they are regarded as common knowledge. This obviously needs to be fixed in the future.

From now on we often suppress consideration of tense and time in expressing ATT-Meta hypotheses, and express them instead in the present tense, as ATT-Meta currently has no treatment of time. We will also often express the hypotheses without using pronouns, as ATT-Meta has no general pronoun-like facility.

10 Conflict Resolution and Specificity

When reasoning is uncertain, examination of a hypothesis H must often involve looking at evidence against as well as for H, in other words looking at evidence for NOT(H). Naturally, there are exceptions. One very important exception is when H is only of interest if it can be established to at least a particular level δ of certainty, because it is used merely as a premise for some argument, not as an end in itself. If even the evidence *for* H cannot get H up to that level, there is no point looking at evidence for NOT(H).

When H and NOT(H) are both strongly supported there is a *conflict-resolution* problem. The present section examines aspects of this problem. One important way conflict can arise is when there is an exception to a default. For instance, as Allen (p.c.) points out, at a stop sign [in the U.S.] one should normally stop. So, in the absence of special circumstances there is simply strong evidence for the proposition that one should stop. However, in special circumstances there can be strong evidence that one should not stop. One then has the problem of trying to decide which body of evidence wins: the general rule or the exceptional situation. Another way that exceptions can arise is in the form of "presumptions" in the legal sense (see Allen, this volume). Conflict can also arise out of inconsistent primary data (as when one witness says the light was red and another that the light was green – Allen, this volume) or from conflicting sets of rules working on the same data.

It is common in AI to propose resolving conflicts by examining the evidential bases for H and NOT(H) and seeing whether one basis is stronger than the other, where the notion of strength is based at least in part on the relative *specificity* of the two evidential bases. A basic example of this is when H is the hypothesis that Peter can fly, based on the fact that Peter is a bird, and NOT(H) is based on the fact that Peter is a penguin. To say that something is a penguin is to be more specific than to say it is a bird, so NOT(H) wins. Clearly, things can get much more complex than in this simple example, because H and NOT(H) may be more distantly related to the facts supporting them, there may be several arguments for H or NOT(H), and each argument may use several facts. A simple form of specificity reasoning is historically central in the sub-areas of AI and cognitive

psychology concerned with semantic networks (see Findler 1979, Lehmann 1992), a knowledge representation tool that relies on net structure to encode what is more specific than what. There is a considerable literature on more advanced forms of specificity (e.g., Loui 1987, Loui *et al.* 1993, Poole 1991, Yen *et al.* 1991, Delgrande & Schaub 1994, Hunter 1994) but existing schemes are far from being the last word on the subject.

Notice that Loui *et al.* (1993) discuss specificity in the domain of legal reasoning. Specificity is indeed a natural thing to include in an AI system for law or ethics, since so much in legal proceedings and ethical judgments consists in showing that some situation is exceptional with regard to some default principle or past general practice. But specificity is not the only tool that has been explored for resolving conflicts. Another one is the use of rule priorities or explicit knowledge about how individual rules defeat each other, as in Hage (1995) in the case of legal reasoning. See also the work of McLaren & Ashley (1995) on reasoning about reasons, with application to law.

ATT-Meta tries to resolve conflicts using specificity, mostly. The basic intuition behind ATT-Meta's specificity comparison algorithm is that if one of the competing hypotheses needs more hypotheses to support it than the other competing hypothesis does, then the former competitor wins. The details of the application of this principle are complex, and are not important for the concerns of this paper. A definition of an early version of ATT-Meta's specificity-comparison algorithm can be found in Barnden *et al.* (1994), but the details have changed greatly since then, and are subject to further occasional change.

What is important for present purposes is what happens when ATT-Meta does or does not find a difference of specificity between conflicting hypotheses H and NOT(H). These are only in conflict if the evidence contributes a rating of `presumed` to each one. When ATT-Meta can determine that the evidence for one of H and NOT(H) is more specific than that for the other, then it downgrades the losing hypothesis to a lower level of certainty and maintains the `presumed` level of the other. On the other hand, if ATT-Meta fails to find a specificity difference, it downgrades both hypotheses. A downgrade reduces the hypothesis's certainty to `suggested` if it has any supporting rule-applications that yield at least `suggested`. Otherwise, the certainty is downgraded to `possible`.

11 Conflict All the Way Down

Conflict-resolution is needed within simulations as well as at the top layer of reasoning. This raises the difficult question of how to coordinate conflict-resolution on different layers. (Analogous questions arise for competing, entirely non-simulative accounts of belief reasoning — the complexities to be discussed betray an inherent complexity in the class of reasoning scenarios considered, not of simulative reasoning in itself.)

Suppose X is (allegedly) simulating Y, and is investigating Y-believes-R, NOT(Y-believes-R), Y-believes-NOT(R), NOT(Y-believes-NOT(R)). X could be the sys-

tem itself, or some agent at an intermediate layer in a tower of simulation layers. Within the simulation of Y, the relevant hypotheses are R and NOT(R). As explained above, negations are not always considered, but for simplicity I will assume here that we are in a situation where each negation just mentioned does need to be considered.

There can be rule applications supporting hypotheses both within in the Y layer and within the X layer. Below I will assume that the rule applications for each of the hypotheses taken individually would give that hypothesis a rating of `presumed` were it not for conflict with negations. The general situation is depicted in Figure 1.

FIGURE 1 ABOUT HERE

A concrete example might be a murder scenario. R is the proposition that Roger is the murderer, and the beliefs of Y appear to support that proposition more strongly than its negation. So far, it seems as though X should conclude that Y believes R. However, Y, who is regarded as an honest person, has apparently sincerely told various people that that Roger is not the murderer, so there is strong non-simulative support for the proposition that she believes that Roger is not the murderer, i.e. Y-believes-NOT(R). There might also be evidence for the negations of the two belief hypotheses. In this example the "direction" of some evidence in the X layer is opposite to that in the Y layer. An alternative possibility, of course, is that the direction of the evidence is the same in both layers. This would be the case if Y has apparently sincerely stated that Roger is the murderer.

Technically, a battle between Y-believes-R and NOT(Y-believes-R) is separate from a battle between Y-believes-NOT(R) and NOT(Y-believes-NOT(R)). However, ATT-Meta has a rule that says that if Y believes something Q then, presumably, Y lacks the belief that NOT(Q). This rule is relevant here, with Q being R or NOT(R). So the two battles actually share whatever evidence there is for Y-believes-R and Y-believes-NOT(R). In the battle between Y-believes-R and its negation, the evidence for the latter includes that for Y-believes-NOT(R), and similarly for the other battle.

The simplest type of case is where, actually, there is no evidence within the Y layer supporting either R or NOT(R). Then, conflict resolution can proceed in the normal way between Y-believes-R and NOT(Y-believes-R), and also between Y-believes-NOT(R) and NOT(Y-believes-NOT(R)).

The opposite type of case is when in the X layer there is no evidence for or against the hypotheses that Y-believes-R and Y-believes-NOT(R). This is depicted in Figure 2. Unfortunately, this does not mean that conflict-resolution can necessarily be done in the Y layer. It may be that the support for R or NOT(R) includes a hypothesis S where one or more of the corresponding X-layer hypotheses (Y-believes-S, and so on) do have evidence for them in the X layer. Therefore this non-simulative evidence is, via the simulated reasoning connecting

S to R, evidence also for one or more of Y-believes-R and so on. As a result, the situation is as complex as when there is *direct* non-simulative evidence for one or more of Y-believes-R and so on.

FIGURE 2 ABOUT HERE

We now proceed to describe the general approach that ATT-Meta takes, and that copes with all the types of situations we have considered. Basically, ATT-Meta "lifts" the evidence in the Y layer up into the X layer, in preparation for *possibly* doing all the conflict-resolution in the X layer. The lifted evidence is used in an algorithm that decides whether to do conflict-resolution in the X layer, or whether it should be done in the Y layer. (Of course, different decisions will in general be made for different hypotheses R.) We will first look at lifting itself and then explain its role in conflict resolution.

11.1 Rule Lifting

Whenever there is a successful rule application in any simulation, the application is "lifted" into the layer just above. That is, if a rule A supports R within the simulation of Y, a lifted form of the rule application is attached to the hypothesis in the X layer that Y-believes-R. The idea is straightforward and can be illustrated with a simple example, depicted in Figure 3. Let Y be Vic in the mugging example, R the hypothesis that Perp hurts Vic, and P the hypothesis that Perp mugs Vic. Suppose there is the following default rule, named "Mugging-Injury":

IF person M mugs person V

AND M is violent

THEN PRESUMABLY M hurts V

Assume that (according to X) Y believes that Perp mugs Vic, and Y believes that Perp is violent. Thus, within the Y layer we have the hypothesis that (a) Perp mugs Vic and the hypothesis that (b) Perp is violent. The hypothesis that Perp hurts Vic is annotated with a *rule-application record* that includes the name Mugging-Injury and a list of the supporting hypotheses (a) and (b). Accordingly, the X-layer hypothesis that Y believes that Perp hurts Vic is annotated with a "lifted rule application" named Λ(Mugging-Injury) and lifted versions of hypotheses (a) and (b), namely Y-believes-(a) and Y-believes-(b).[8]

FIGURE 3 ABOUT HERE

The reader will have noticed the extra hypothesis in the lifted rule application in the figure. We call such hypotheses *agent-inference hypotheses.* Recall that X is merely alleging that Y does the inference from (a) and (b) to R. But something may, in fact, prevent Y from doing it. For instance, Y may fail to consider (a) and

(b) together. So, ATT-Meta sets up the explicit hypothesis, in the X layer, saying that Y makes inferences from (a) and (b).[9] This hypothesis can be reasoned about just like any other. However, in keeping with the nature of simulative reasoning, the system has a built-in rule that says that any given agent-inference hypothesis is presumably true. This rule is rather special in not relying on any evidence, so that in the X/Y scenario if there *is* any evidence that Y presumably did *not* make inferences from (a) and (b), then this evidence prevails and the agent-inference hypothesis is defeated (i.e., downgraded in certainty).

Now recall that each rule has its own confidence qualifier (C_0 above), used as a limit on the confidence level that the rule can return. The rule's qualifier is included in any rule-application record resulting from that rule. Lifted rule-applications also include a qualifier. It is always `presumed.` This means that, even when it is *certain* that Y believes (a), that Y believes (b), and that Y does inferences from (a) and (b), the lifted application contributes at most a `presumed` rating to Y-believes-R. This is because the fact that Y applies a rule in support of R still leaves open the possibility that Y fails to believe R because of other effects (e.g., having a counter-argument to R that is unknown to X).

Finally, lifted rule applications are themselves subject to lifting just as any ordinary application is. Thus, if X is being simulated by another agent W, the lifted applications in the X layer resulting from rule-applications within the Y-layer are themselves lifted to W's own layer.

11.2 Rule Lifting and Conflict Resolution

In the difficult cases of multi-layer conflict resolution, there are relevant rule applications both within the X layer and the Y layer. However, the basic principle is that *the applications within the Y layer are lifted into the X layer, so that they can take part in conflict resolution in concert with ordinary applications in the X layer, using the ordinary specificity-comparison algorithm mentioned in section 10.* For example, in the Perp-mugging-Vic case, we get the lifted application like the one depicted in Figure 3 supporting Y-believes-R, but we may also have, say, an ordinary rule application supporting Y-believes-NOT(R).

Assuming that the conflict between competing hypotheses Y-believes-R and NOT(Y-believes-R) is being considered in the X layer, the following occurs. The overall strategy is to see whether conflict-resolution between those hypotheses needs to be done in the X layer, or whether instead the system should look inside the Y simulation and consider the conflict between R and NOT(R). ATT-Meta does conflict-resolution in the X layer between Y-believes-R and NOT(Y-believes-R) if and only if at least one of these hypotheses has "unlowerable support." A hypothesis H has unlowerable support if, roughly speaking, some non-"finalized" hypothesis directly *or indirectly* supporting H (including H itself) has a non-lifted rule-application supporting it. A hypothesis is finalized if all decisions, including ones emanating from conflict-resolution, have been done for it. The process by which finalizedness is determined will not be detailed here.

If at least one of Y-believes-R nor NOT(Y-believes-R) has unlowerable support, ATT-Meta applies the ordinary specificity comparison algorithm alluded to in section 10 to these contending hypotheses – *it involves no special treatment of lifted applications or hypotheses about belief.* As normal, if a winner is found, the other contender is downgraded; if not, both contenders are downgraded. A similar process occurs with Y-believes-NOT(R) and NOT(Y-believes-NOT(R)). Downgrade of, say, Y-believes-R causes the rule applications supporting R within the Y simulation to be suppressed, causing R itself to go down in certainty, so that there is no need for conflict resolution between R and NOT(R). (Special action is taken in the rare cases when Y-believes-R and Y-believes-NOT(R) both win their battles.)

If, on the other hand, neither Y-believes-R nor NOT(Y-believes-R) has unlowerable support, those hypotheses are *not* downgraded, and instead ATT-Meta will descend into the Y simulation and consider the conflict between R and NOT(R). The resolution of this conflict will then indirectly cause the conflicts in the X layer to be resolved.

We have only been considering two layers, the X layer and the Y layer. But these could be intermediate in a tower of layers: for instance, R could be about a belief of some agent, and/or the X layer may be a simulation layer within another layer. ATT-Meta proceeds by considering conflicts in the top layer, resolving them when appropriate and possible, then moving down to the next layer, and so on.

12 ATT-Meta and the Mugging Example

Here we explain how ATT-Meta can make some of the inferences discussed in the mugging example in section 2. A particular reason for treating the example is to show how conflict-resolution occurs and (in a small way) how simulative and non-simulative reasoning can mix. The inferences made in this section are not earth-shattering in importance, but the processes involved are indicative of what is needed also in much more portentous cases. The reader should be aware that no attempt is made in the example to include the real legal considerations that might arise in a mugging case. Also, to make the presentation manageable, little mention will be made of the certainty levels entertained by agents in their beliefs.

The rules used in this section and elsewhere in the paper are merely illustrative, and are in general simpler, and much less numerous, than the rules that would be needed in a real application of ATT-Meta to the legal domain. Also, the example involves the connection between what people say and what they believe, conditions under which people might lie or be mistaken, and people's beliefs about actions impinging upon them. These are of course very complex matters, so the present section includes only highly simplified, skeleton accounts, taking enormous shortcuts through the real complexities. As a result, the reasoning rules included are necessarily tailored to the needs of the specific example.

Sections 12.1 and 12.2 deal with different versions of the example. Over the two runs of ATT-Meta for these versions, the system exhibited a maximum hypothesis explosion factor (see section 5) of 1.36, without taking into account agent-inference hypotheses, and of 1.58 when all hypotheses were considered. The system created a maximum of 50 hypotheses, although many of these led to little processing as they received no support at all. We will discuss only the most important hypotheses generated.

12.1 Showing that Vic believes Perp Not Malicious

First, let us see how ATT-Meta infers that, presumably, Vic believes that Perp is not trying to hurt him during the mugging. The main hypotheses are as shown in Figure 4.

FIGURE 4 ABOUT HERE

We assume that ATT-Meta is given the following specific facts about the situation, where all these facts have certainty level `certain`:

- Vic says that Perp mugs Vic.
- Vic says that Perp does not try to hurt Vic while mugging Vic.
- Vic says that Vic does not know Perp.
- X mugging Y is a direct physical act of X upon Y.

We assume that ATT-Meta has the following rules:

(a) IF person M mugs person V
THEN PRESUMABLY M tries to hurt V while mugging V

(b) IF person Y says that Q
THEN PRESUMABLY Y believes that Q

(c) IF person Y says that person X performs action A on Y
AND A is a direct physical act of X on Y
AND Y says that Y does not know X
THEN PRESUMABLY Y believes that X performs action A on Y.

(d) IF person Y believes that X performs action A on Y
AND Y says that X does not do B while doing A
AND A is a direct physical act of X on Y
AND Y says that Y does not know X
THEN PRESUMABLY Y believes that X does not do B while doing A.

Notice that the certainty level of each rule is `presumed`, so conclusions from the rules can never get above `presumed` unless supported by other means.

Rules (c) and (d) advert to more specific situations than (b) does, even though they all conclude that someone believes something from what they say. Rule (b) is a basic default, and its conclusions could be contradicted by rules that consider specific situations, such as a rule that said that if someone says something and is lying then they do not believe what they say. Rules (c) and (d) are meant to partially capture that default that people are generally honest when describing physical actions they think other people perform on them, when they do not know who those people are (so they have no axes to grind). Rule (d) includes relativization to an action A to avoid inclusion in this example of a rich theory of actions and time, and of statements about them (ATT-Meta currently has no treatment of time). Observe that Vic's statement to the police that "Perp was not trying to hurt me" should be understood as conveying "during the mugging" or "while mugging me."

The user of the system creates the top goal hypothesis (labeled 1- in the figure). It comes to be supported by an application of rule (d) to the hypotheses labeled (4),(6),(7) and (8) in the figure, and also by an application of rule (b) to fact (7). Because of the creation of the top goal, there is a simulation of Vic, and inside it is placed the lowered hypothesis, namely that Perp does not try to hurt Vic (-2) while mugging him. Now, because the top goal (1-) has high-certainty support, its negation (-1-) is also created for investigation. This then leads to the creation of the hypothesis (1+) that Vic believes that Perp tries to hurt Vic while mugging him, for investigation.

The creation of this node is accompanied by the creation of the lowered hypothesis (2) [Perp tries to hurt Vic while mugging him] inside the simulation. This hypothesis is supported by an application of rule (a) to the within-simulation hypothesis that Perp mugs Vic (3).

The creation of the rule application linking (3) to (2) inside the simulation is accompanied by the creation of the lifted rule application linking (4) to (1+). Hypothesis (4) [Vic believes that Perp mugs him] is supported by an application of rule (c) to facts (5, 6, 8), and also by an application of rule (b) to fact (5). Thus, hypothesis (1+) gets a `presumed` certainty-level contribution from its support, and therefore so does (-1-), which is the negation of the top goal. We therefore have two conflicts outside the simulation: between (1+) and (-1+), and between (1-) and (-1-).

At a certain point facts (5), (6), (7) and (8) become finalized, and this then allows hypothesis (4) [Vic believes that Perp mugs Vic] to be finalized. Let us look at the conflict between (1-) and (-1-). Even though the latter does not have unlowerable support, because (4) is finalized, (1-) is not yet finalized and is supported by a non-lifted rule application, and therefore has unlowerable support. So ATT-Meta applies conflict resolution to (1-) and (-1-). The same thing happens in the case of the conflict between (1+) and (-1+).

Specificity comparison between (1-) and (-1-) decides in favour of the former, essentially because one argument for the former needs all four facts (5,6,7,8)—

the need for (5) being indirect via (4)—whereas the arguments for (-1-) only need facts (5,6,8)—indirectly via (1+) and (4). Thus, (1-) arises from a more specific situation than (-1-) does. By a similar analysis, specificity comparison decides in favour of (-1+) over (1+).[10]

As a result, the losing hypotheses (1+ and -1-) are downgraded to `suggested`. The downgrade of (1+) causes (2) [Perp tries to hurt Vic while mugging V] inside the simulation to be capped at `suggested`. As a result, it is (-2) [Perp does not try to hurt Vic while mugging Vic] that survives as `presumed` within the simulation.

In effect, the conflict within the simulation is resolved as in indirect result of the resolution of the conflicts outside the simulation.

12.2 Remarks on a Variant Scenario

Suppose now that instead of Vic saying that Perp did not try to hurt him while mugging him, Vic says that Perp is non-violent. See fact (7′) in Figure 5. Suppose also there is a rule that says that

(e) IF someone X is non-violent

AND X performs action A on Y

THEN PRESUMABLY X does not try to hurt Y while performing A.[11]

We then have new substructure inside the simulation, as shown in the figure. Once hypotheses (4) [Vic believes that Perp mugs Vic] and (9) [Vic believes that Perp is non-violent] are finalized, it transpires that none of (1-), (-1-), (1+), (-1+) have unlowerable support. Therefore, in this variant scenario ATT-Meta does not do conflict resolution outside the simulation, and instead descends into the simulation to consider the conflict between (2) and (-2) directly.

FIGURE 5 ABOUT HERE

Hypothesis (-2) [Perp does not try to hurt Vic while mugging Vic] wins by the basic specificity-comparison algorithm. This is because (-2) needs both (3) and (10), whereas (2) only needs (3). Because (-2) wins, its negation (2) is downgraded to `suggested`. This causes the rule-application supporting (2) to be "suppressed," and this then causes the lifted version of the application, joining (2) to (1+), also to be suppressed. As a result, the certainty level of (1+) [Vic believes Perp tries to hurt him while mugging him] is reduced (to `suggested`). This causes the certainty level of (-1-), which is supported only by (1+), to go down to `suggested`. On the other hand, there is no such effect on node (1-), so this hypothesis implicitly wins over its negation(-1-) as an indirect result of the conflict-resolution inside the simulation. Similarly, (-1+) indirectly wins over (1+).

12.3 Showing Perp Not Malicious

Let us now see how ATT-Meta would infer that, presumably, Perp does not try to hurt Vic while mugging him. We will assume that ATT-Meta has the facts and rules listed above, together with the following fact:

X trying to hurt Y is a direct physical act of X upon Y

shown as (8′) in Figure 6, and the following rules:

(f) IF A is a direct physical act upon Y
AND Y believes that A happens
THEN PRESUMABLY action A happens

(g) IF A is a direct physical act upon Y
AND NOT(Y believes that A happens)
THEN PRESUMABLY NOT(action A happens).

FIGURE 6 ABOUT HERE

We suppose that the user creates the top goal, namely that Perp does not try to hurt Vic, shown as (-0) in Figure 6. This, partly via the whole process described in section 12.1 or 12.2, ultimately gets supported by an application of rule (g) acting on hypotheses (8′) and (-1+). These hypotheses are that Perp's trying to hurt Vic is a direct physical act by Perp on Vic, and the hypothesis that Vic lacks the belief that Perp tries to hurt him while mugging him. Recall that this hypothesis keeps a high level of certainty as a result of the conflict resolution in the process described in subsection 12.1 or 12.2, and that it is supported by (1-) [Vic believes that Perp does not try to hurt him ...].

However, the negation (0) of the top goal also gets strong support by an application of rule (a) to the hypothesis (11) that Perp mugs Vic. This hypothesis gets strong support from an application of rule (f) to the fact (8) that Perp mugging Vic is a direct physical act of Perp on Vic and the hypothesis (4) that Vic believes that Perp mugs him.

We therefore have a conflict between the goal and its negation. The facts needed by the neediest argument for the goal are (5,6,7,8,8′), because all these except (8′) are needed by an argument for (1-) and hence by (-1+). But the only facts needed by the arguments for the goal's negation (0) are (5,6,8). As with conflicts in the previous subsections, it is simple for the specificity-based conflict-resolution mechanism to decide that the goal is more specifically supported and should therefore win the conflict.

12.4 Embedding within Another Agent

The whole of the inferencing so far described can be embedded inside another agent. For example, ATT-Meta could reason about a juror doing the reason portrayed above. That reasoning would now be in a simulation of the juror, and the simulation of Vic would be nested inside it.

All hypotheses H outside Vic in Figures 4 (or 5) and 6 are accompanied outside the juror by hypotheses of the rough form "juror believes that H." It will then turn out that there are conflicts between belief hypotheses outside the juror. However, assuming that there is no special knowledge about what the jurors believe, other than that they believe the facts listed above, it will turn out that conflict resolution will not be done outside the juror simulation, but will instead be done inside. The conflict resolution between the juror's hypotheses about Vic's beliefs will then be done outside the juror's Vic simulation as in section 12.1, or within that simulation as in 12.2.

12.5 Vic's Actual Perspective

A motive for using a real mugging event that occurred to the present author was that it enables us realistically to consider both the actual inner perspective of Vic as well as the perspective of people thinking about the scenario from the outside. I know that I came to the conclusion that the muggers were not trying to hurt me, contradicting my default that muggers try to hurt their victims. Thus, I in fact did some conflict-resolution in some way. However, under some circumstances an observer of the situation using a layered conflict-resolution regime like the one above would do conflict resolution *outside* their simulation of me, as we saw in section 12.1 and Figure 4. When the result of doing that is to ascribe to me the belief that the muggers were not trying to hurt me, we can say that my conflict-resolution has in the observer's deliberations been lifted outside me. In a case where the outside conflict resolution does not ascribe to me the belief that the muggers were not trying to hurt me (and perhaps to ascribe to me the belief that the muggers *were* trying to hurt me), we can say that the observer has omitted to ascribe to me the conflict-resolution that I in fact did. However, that is not a deficiency of the above methods — it merely reflects the fact that the observer does not have information to be able to defeat the evidence that leads that observer to fail to lift my conflict resolution.

13 Conclusion

We have seen how questions of uncertainty are important in the application of belief reasoning to the legal domain, and how they greatly complicate the conduct of belief reasoning. Uncertainty is inherent in important inference techniques such as default reasoning, abduction and case-based/analogy-based reasoning, and agents that are being reasoned about may use these techniques; also, any

reasoning about people's mental states or processes is fraught with uncertainty since we do not in fact have access to their minds. The two types of uncertainty can crop up at any layer of reasoning within a nested-belief situation.

Although the focus has been on simulative reasoning, it should not be thought that the complications are the fault of simulative reasoning. For instance, the explosion of hypotheses in Section 5 obtains whether or not the belief-reasoning is simulative, and the need to coordinate conflict resolution at different layers of belief arises whether or not simulation is used. Coping with conflict resolution in an entirely non-simulative system would be more complex than in a simulative one, in fact, because one would need some way of externally modelling the conflict resolution acts of agents that are being reasoned about.

We have presented the ATT-Meta approach to some of the issues raised. However, the point of doing this was merely to clarify the nature of the issues and to suggest that the issues can be addressed satisfactorily, rather than give a complete description of ATT-Meta or to propose ATT-Meta as a complete solution to the problems. It contains, nevertheless, one of the most complete integrations of uncertainty-handling and belief reasoning currently available in AI.

The comments in this paper about ATT-Meta are complex enough as they are, but in fact they are considerably over-simplified compared to the reality of the system. For instance, the system maintains several ancillary certainty measures for each hypothesis, not just a single certainty level per hypothesis. Also, the handling of rules is more complicated than I have portrayed; for instance, the application of a rule can be suspended half-way through if it looks unpromising, and then resumed later when new information is obtained. Also, there are major complications involved in deciding when exactly during processing to try to resolve conflicts, a matter that is exacerbated by the high degree of cyclicity that arises in practice within the network of interdependencies between hypotheses.

ATT-Meta has some major lacunae. There are some technical restrictions in the reasoning it can do. For example, it does not currently have a full treatment of quantification; and it cannot reason in a case-by-case way. (I am not alluding to case-based reasoning here. Rather, case-by-case reasoning is follows: conclude R from P-or-Q, given that R follows from each of P and Q individually. ATT-Meta cannot currently do this unless some rule mentions P-or-Q explicitly.) ATT-Meta is virtually devoid of reasoning about time and change, although I recognize that these matters are important for mental state reasoning, not least in its application to law.

Acknowledgment

This work was supported in part by grants IRI-9101354 and CDA-8914670 from the National Science Foundation of the USA.

Notes

(1) We will frequently be making reference to Allen (this volume) and Allen (personal communication), abbreviated to Allen (p.c.), because he has provided scenarios and arguments that provide useful background for the present paper.

(2) I exclude here most work on "truth maintenance" and "belief revision" (see Rich & Knight 1991 for an introduction), because the research is on management of a system's own beliefs, rather than on reasoning about beliefs of other agents. For the same reason, I also exclude the large amount of work done in recent years on "belief networks" (e.g., Pearl 1986). These, again, are for computing probabilities of propositions entertained by the reasoning system itself.

(3) "ATT-Meta" stands for "[propositional] attitudes" and "metaphor." There is no connection to the AT&T company! ATT-Meta's capability for metaphor-based reasoning is not described here—see Barnden *et al.* (1994, 1996) and Barnden (1998a,b).

(4) For brevity, I often use a neuter word when referring to an agent if it could be non-human, even if it could alternatively be human, to avoid devices such as "it/she/he."

(5) I include case-based reasoning within analogy-based reasoning, since analogy between cases is the central idea of case-based reasoning.

(6) Barnden (1997) is a preliminary version of the current article. Some of the details in it concerning conflict-resolution are out of date, but the information on optimization is still valid.

(7) In ATT-Meta, a working assumption is a hypothesis that has the certainty-level called "`presumed`". This is discussed further in section 9.

(8) This discussion suppresses the question of Y's level of confidence within hypotheses such as Y-believes-R. Lifting is only done if the base rule, e.g. Mugging-Injury, yields a level of confidence at least as high as the one mentioned in the Y-believes-R hypothesis.

(9) This falls short of what is really required, as it does not specify the nature of the inference more tightly. This deficiency will be corrected in future versions of ATT-Meta.

(10) Specificity-comparison does not just look at facts, but I concentrate on just this aspect for simplicity of presentation. Note also that hypotheses (4) and (-1) are supported in a relatively unspecific way by applications of rule (b), as well as by the applications of rules (c) and (d). The (b) applications do not upset the specificity comparison.

(11) This rule would benefit from having "AND A is a direct physical act of X on Y" in its IF part, but we omit this for simplicity of illustration. We would need an extra rule to establish that Vic believes that mugging is a direct physical act.

References

Asher, N. & Lascarides, A. (1994). Intentions and information in discourse. In *Procs. 32nd Annual Meeting of the Association for Computational Linguistics,* pp.34–41. Association for Computational Linguistics.

Attardi, G. & Simi, M. (1994). Proofs in context. In J. Doyle, E. Sandewall & P. Torasso (Eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference,* pp.15–26. (Bonn, Germany, 24–27 May 1994.) San Mateo, CA: Morgan Kaufmann.

Ballim, A. & Wilks, Y. (1991). *Artificial believers: The ascription of belief.* Hillsdale, N.J.: Lawrence Erlbaum.

Barnden, J.A. (1995). Simulative reasoning, common-sense psychology and artificial intelligence. In M. Davies & T. Stone (Eds), *Mental Simulation: Evaluations and Applications*, pp.247–273. Oxford, U.K.: Blackwell.

Barnden, J.A. (1997). Simulation and uncertainty in reasoning about agents' beliefs. *Memoranda in Computer and Cognitive Science*, No. MCCS–97–310, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003, U.S.A.

Barnden, J.A. (1998a). An AI system for metaphorical reasoning about mental states in discourse. In Koenig, J-P. (Ed.), *Discourse and Cognition: Bridging the Gap.* Stanford, CA: CSLI.

Barnden, J.A. (1998b). Combining uncertain belief reasoning and uncertain metaphor-based reasoning. In *Procs. Twentieth Annual Meeting of the Cognitive Science Society,* pp.114–119. Mahwah, N.J.: Lawrence Erlbaum Associates.

Barnden, J.A., Helmreich, S., Iverson, E. & Stein, G.C. (1994). An integrated implementation of simulative, uncertain and metaphorical reasoning about mental states. In J. Doyle, E. Sandewall & P. Torasso (Eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference,* pp.27–38. (Bonn, Germany, 24–27 May 1994.) San Mateo, CA: Morgan Kaufmann.

Barnden, J.A., Helmreich, S., Iverson, E. & Stein, G.C. (1996). Artificial intelligence and metaphors of mind: within-vehicle reasoning and its benefits. *Metaphor and Symbolic Activity, 11*(2), pp.101–123.

Bratman, M.E. (1992). Planning and the stability of intention. *Minds and Machines, 2* (1), pp.1–16.

Carruthers, P. & Smith, P.K. (Eds). (1996). *Theories of theories of mind.* Cambridge, UK: Cambridge University Press.

Chalupsky, H. (1993). Using hypothetical reasoning as a method for belief ascription. *J. Experimental and Theoretical Artificial Intelligence, 5* (2&3), pp.119–133.

Chalupsky, H. (1996). Belief ascription by way of simulative reasoning. Ph.D. Dissertation, Department of Computer Science, State University of New York at Buffalo.

Cravo, M.R. & Martins, J.P. (1993). SNePSwD: A newcomer to the SNePS family. *J. Experimental and Theoretical Artificial Intelligence, 5* (2&3), pp.135–148.

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, CA: Morgan Kaufmann.

Davies, M & Stone, T. (Eds) (1995). *Mental simulation: evaluations and applications*. Oxford, U.K.: Blackwell.

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

Delgrande, J.P. & Schaub, T.H. (1994). A general approach to specificity in default reasoning. In J. Doyle, E. Sandewall & P. Torasso (Eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference,* pp.146–157. (Bonn, Germany, 24–27 May 1994.) San Mateo, CA: Morgan Kaufmann.

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

Dragoni, A.F. & Puliti, P. (1994). Mental states recognition from speech acts through abduction. In *Procs. 11th European Conference on Artificial Intelligence* (ECAI-94), pp.183–187. Chichester, UK: John Wiley.

Findler, N.V. (Ed.) (1979). *Associative networks: Representation and use of knowledge by computers*. New York: Academic Press.

Goldman, A.I. (1992). In defense of the simulation theory. *Mind and Language, 7* (1 & 2), pp.104–119.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language, 1*, pp.158–171.

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

Hage, J. (1995). Teleological reasoning in reason-based logic. In *Fifth International Conference on Artificial Intelligence and Law: Proceedings of the Conference,* p.11–20. New York: Association for Computing Machinery.

Harris, P.L. (1992). From simulation to folk psychology: the case for development. *Mind and Language, 7* (1 & 2), pp.120–144.

Hunter, A. (1994). Defeasible reasoning with structured information. In J. Doyle, E. Sandewall & P. Torasso (Eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference,* pp.281–292. (Bonn, Germany, 24–27 May 1994.) San Mateo, CA: Morgan Kaufmann.

Kaplan, A.N. & Schubert, L.K. (1997a). Simulative inference in a computational model of belief. In H. Bunt, L. Kievit, R. Muskens & M. Verlinden (Eds), IWCS II: *Second International Workshop on Computational Semantics*, Tilburg University, Netherlands, Jan. 8–10, 1997.

Kaplan, A. N. & Schubert, L.K. (1997b). Simulative inference in a computational model of belief. Tech. Rep. 636, Dept. of Computer Science, Univ. of Rochester, Rochester, NY 14627-0226, October 1997.

Konolige, K. (1985). Belief and incompleteness. In *Formal Theories of the Commonsense World,* pp.359–403. Norwood, NJ: Ablex.

Konolige, K. (1986). *A deduction model of belief.* London: Pitman. Los Altos: Morgan Kaufmann.

Konolige, K. (1988). On the relation between default and autoepistemic logic. *Artificial Intelligence, 35* (3), pp.343–382.

Lehmann, F.W. (1992). *Semantic networks in artificial intelligence.* New York: Pergamon. (Special issue of *Computer and Mathematics with Applications, 23,* Nos. 2–9.)

Loui, R.P. (1987). Defeat among arguments: a system of defeasible inference. *Computational Intelligence, 3*, pp.100–106.

Loui, R.P., Norman, J., Olson, J. & Merrill, A. (1993). A design for reasoning with policies, precedents, and rationales. In *Fourth International Conference on Artificial Intelligence and Law: Proceedings of the Conference,* pp.202–211. New York: Association for Computing Machinery.

McLaren, B.M. & Ashley, K.D. (1995). Context sensitive case comparisons in practical ethics: reasoning about reasons. In *Fifth International Conference on Artificial Intelligence and Law: Proceedings of the Conference,* p.316–325. New York: Association for Computing Machinery.

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.

Nissan, E., Puni, G. & Kuflik, T. (1991). Finding excuses with ALIBI: Alternative plans that are deontically more defensible. *Computers and Artificial Intelligence, 10*(4), pp. 297–325. Also in J. Lopes Alves (Ed), *Information Technology & Society: Theory, Uses, Impacts.* Lisbon: Associação Portuguesa para o Desenvolvimento das Comunicações, & Sociedade Portuguesa de Filosofia, pp.484–510 (1992).

Parsons, S., Sierra, C. & Jennings, N. (1998). Multi-context argumentative agents. In *Working Papers of the Fourth Symp. on Logical Formalizations of Commonsense Reasoning* (COMMON SENSE '98), Queen Mary and Westfield College, London, 7–9 Januray 1998.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29* (3), pp.241–288.

Perrault, C.R. (1990). An application of default logic to speech act theory. In P.R. Cohen, J. Morgan & M.E. Pollack (Eds), *Intentions in Communication,* pp.161–185. Cambridge, MA: MIT Press.

Poole, D. (1991). The effect of knowledge on belief: conditioning, specificity and the lottery paradox in default reasoning. *Artificial Intelligence, 49*, pp.281–307.

Rich, E. & Knight, K. (1991). *Artificial intelligence.* 2nd edition. New York: McGraw-Hill.

Russell, S. & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Englewood Cliffs, N.J.: Prentice-Hall.

Yen, J., Neches, R. & MacGregor, R. (1991). CLASP: Integrating term subsumption systems and production systems. *IEEE Trans. on Knowledge and Data Engineering, 3* (1), pp.25–32.

Figure Captions

FIGURE 1: The general case of conflict-resolution in the presence of simulation. Agent X is (allegedly) simulating agent Y. (X itself may be being simulated by an agent further out.) The discs depict hypotheses. A minus sign indicates negation. Negations of hypotheses are shown as shaded discs. The H-shaped connectors emphasize the relationship between complementary hypotheses (hypotheses which are negations of each other). The notation"Y:" means "Y believes that." Questions of uncertainty are suppressed for simplicity of presentation. Dashed lines show correspondences between hypotheses in different layers. Dotted lines show applications of a special belief-management rule. The crooked arrows depict rule-based arguments supporting a hypothesis to level `presumed`.

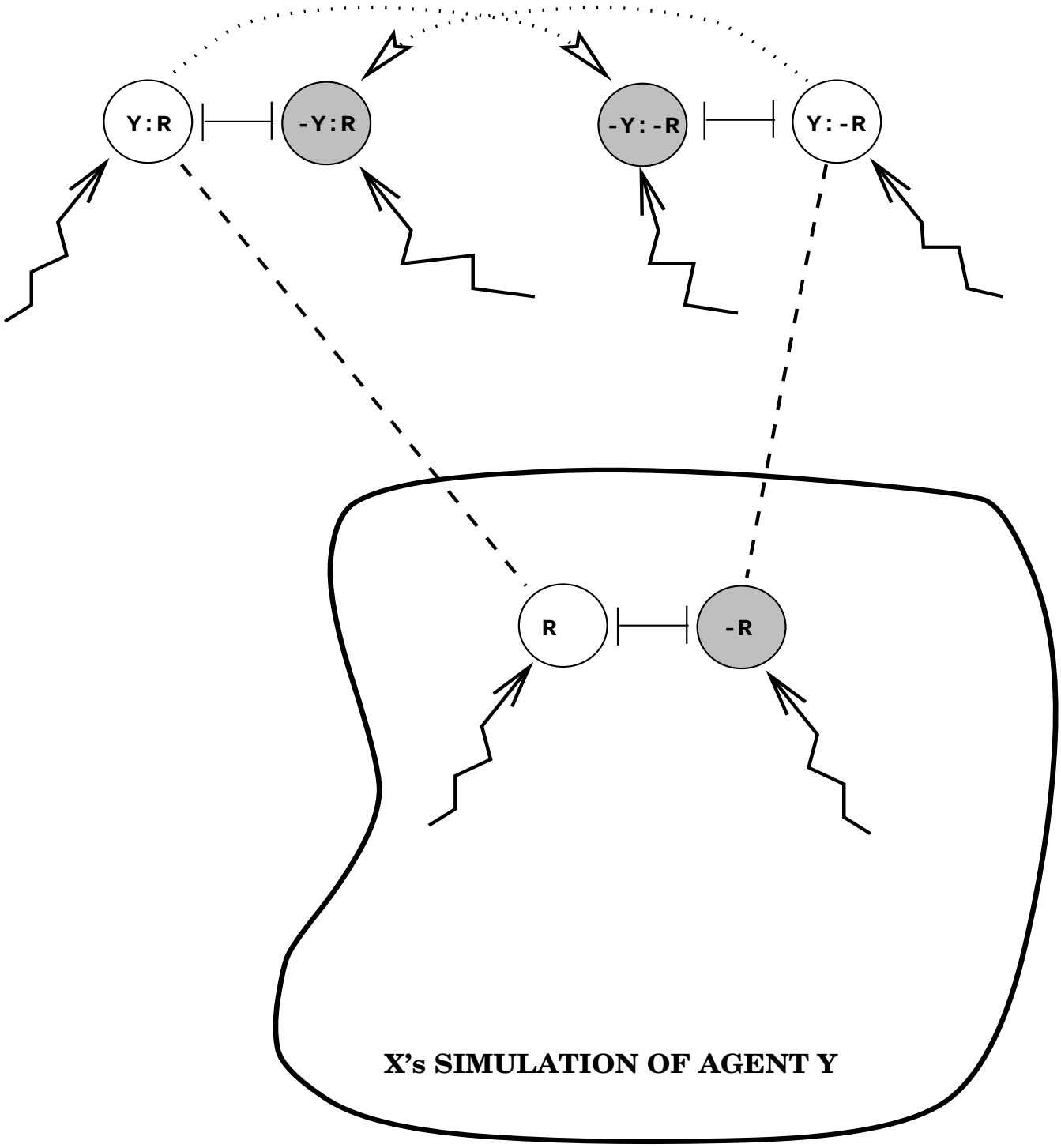
FIGURE 2: A deceptively simple case of conflict-resolution in the presence of simulation. See Figure 1 for diagram conventions. Comparison of specificity of evidence does *not* necessarily occur in the Y layer, because of the possibility of a hypothesis S as shown. S is part of the (direct or indirect) evidence supporting R.

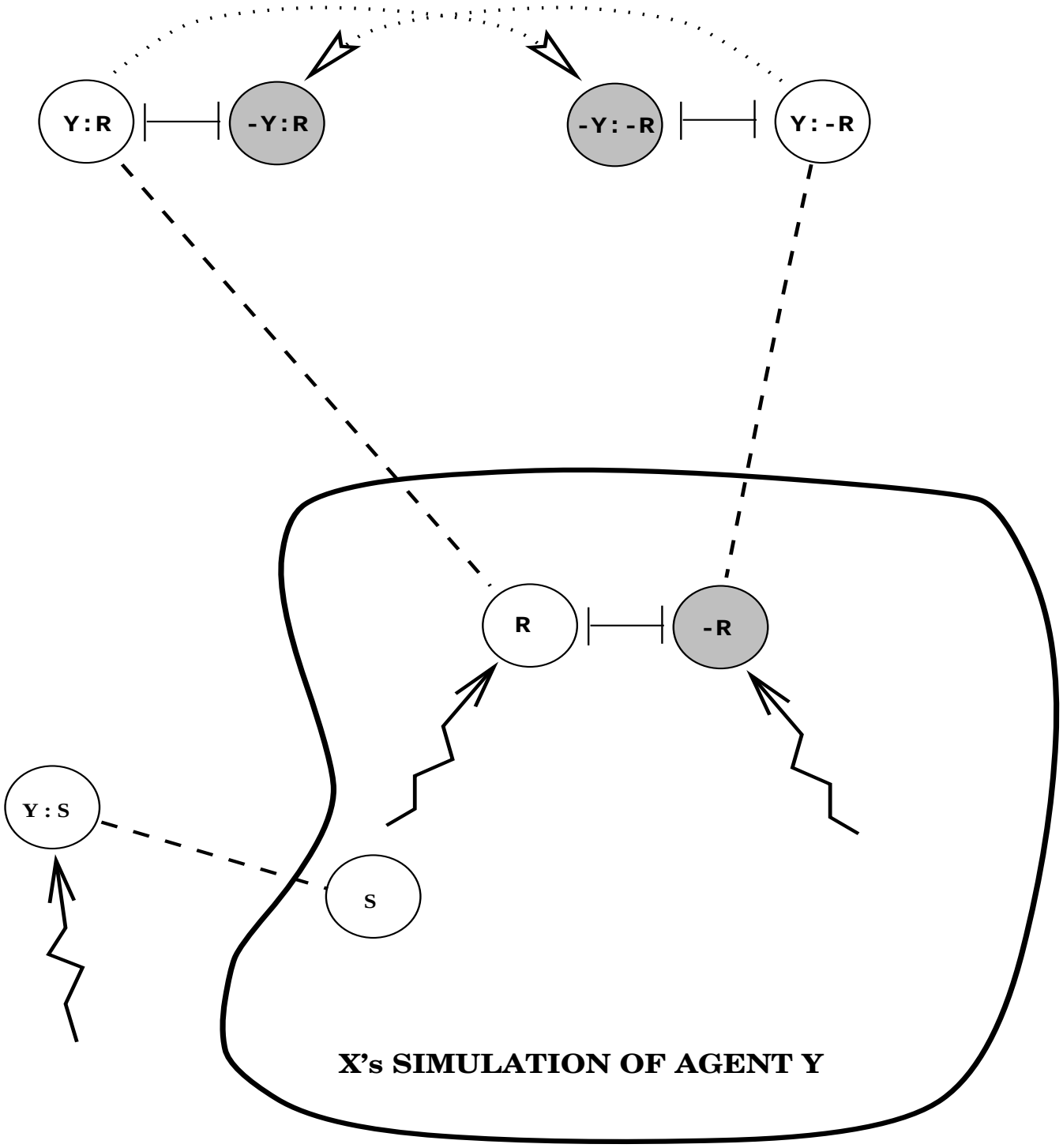
FIGURE 3: Lifting of rule applications in simulations. Within the Y simulation, there is an application of the Mugging-Injury rule, to hypotheses labelled (a) and (b) in the figure. The lifted form of the application has isomorphic structure, except that it has an extra supporting hypothesis. This is an "agent-inference hypothesis" — see text.

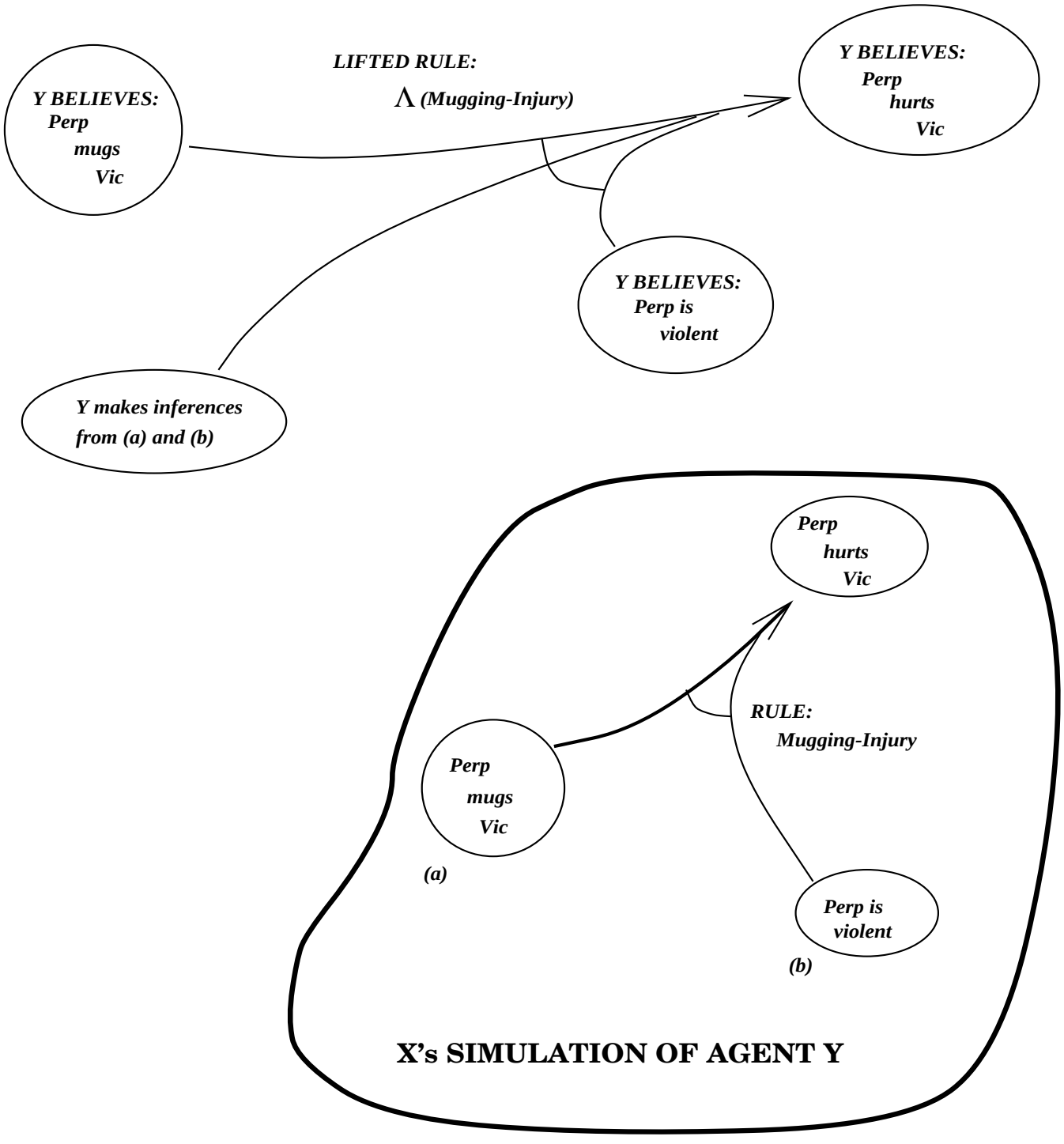
FIGURE 4: Illustrating some of the reasoning performed in ATT-Meta's treatment of the original mugging example, as detailed in section 12.1. P, V stand for Perp, Vic respectively. SAY(h) stands for: Vic says that h. BEL(h) stands for: Vic believes h (to level at least `presumed`). However, **P tryhurt V** is an abbreviation for **P tryhurt V while mugging V**. **dir-phys(mug)** stands for the hypothesis that Perp mugging Vic is a direct physical action by Perp on Vic. A minus sign means negation. Complementary hypotheses are joined by an H-shaped symbol. Dashed lines show correspondences between hypotheses in different layers. Solid arrowed lines show rule applications. A "Λ" indicates a lifted rule application. Dotted lines show applications of a special belief-management rule. Two diagonal lines by a hypothesis mean that the hypothesis is a fact. The numerical symbols next to hypotheses are labels used for reference in the text. The letters next to the rule-application lines identify the rules used. Not all hypotheses that ATT-Meta considers are shown, or discussed in the text.

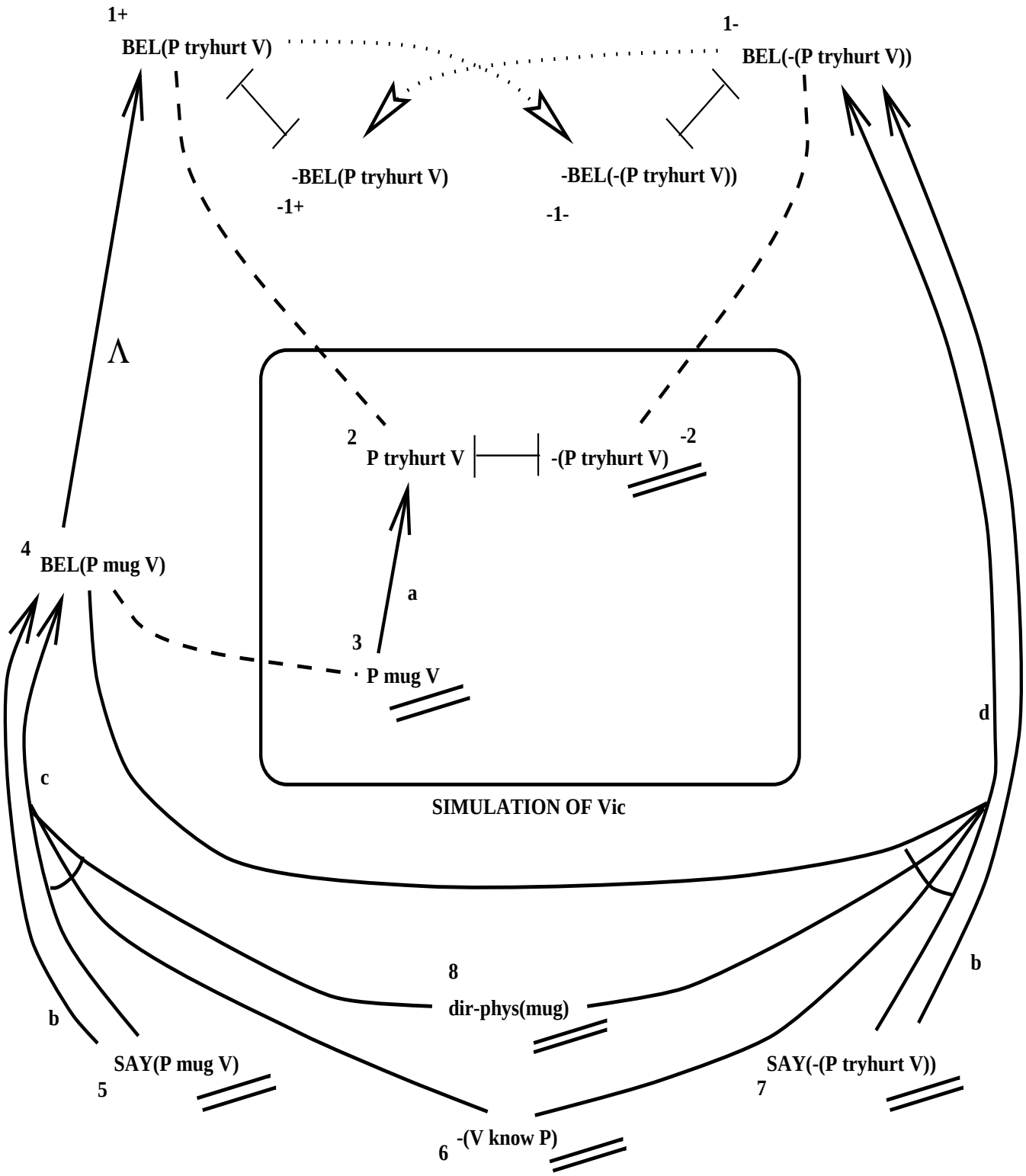
FIGURE 5: Illustrating some of the reasoning performed in ATT-Meta's treatment of the variant mugging example, as detailed in section 12.2. The new hypotheses introduced for this variant are shown surrounded by boxes.

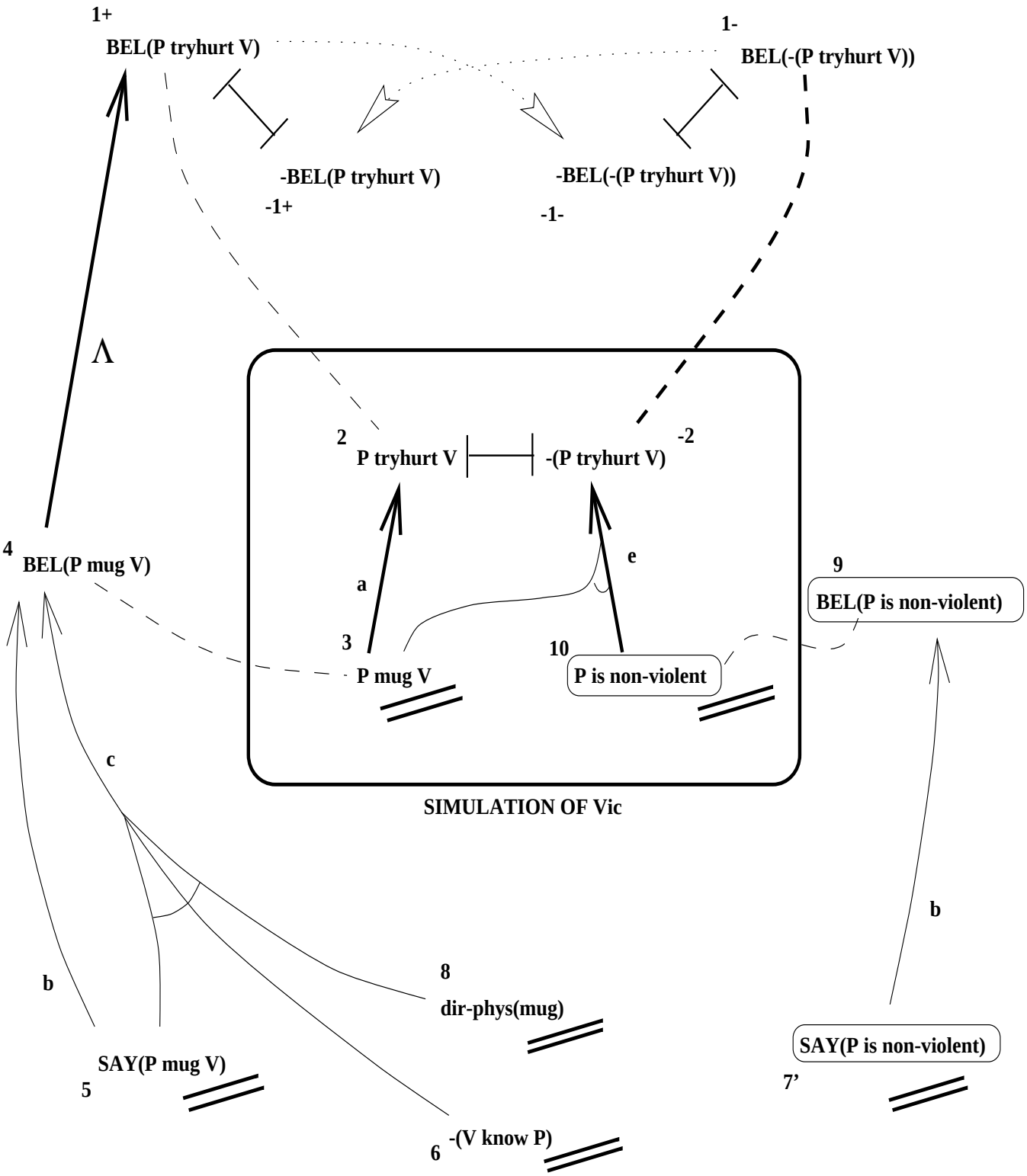
FIGURE 6: Illustrating some more of the reasoning performed in ATT-Meta's treatment of the mugging example, as detailed in section 12.3. The new hypotheses introduced for this explanation are shown surrounded by boxes. The whole of Figure 4 or 5 should be considered to be included in the present figure.

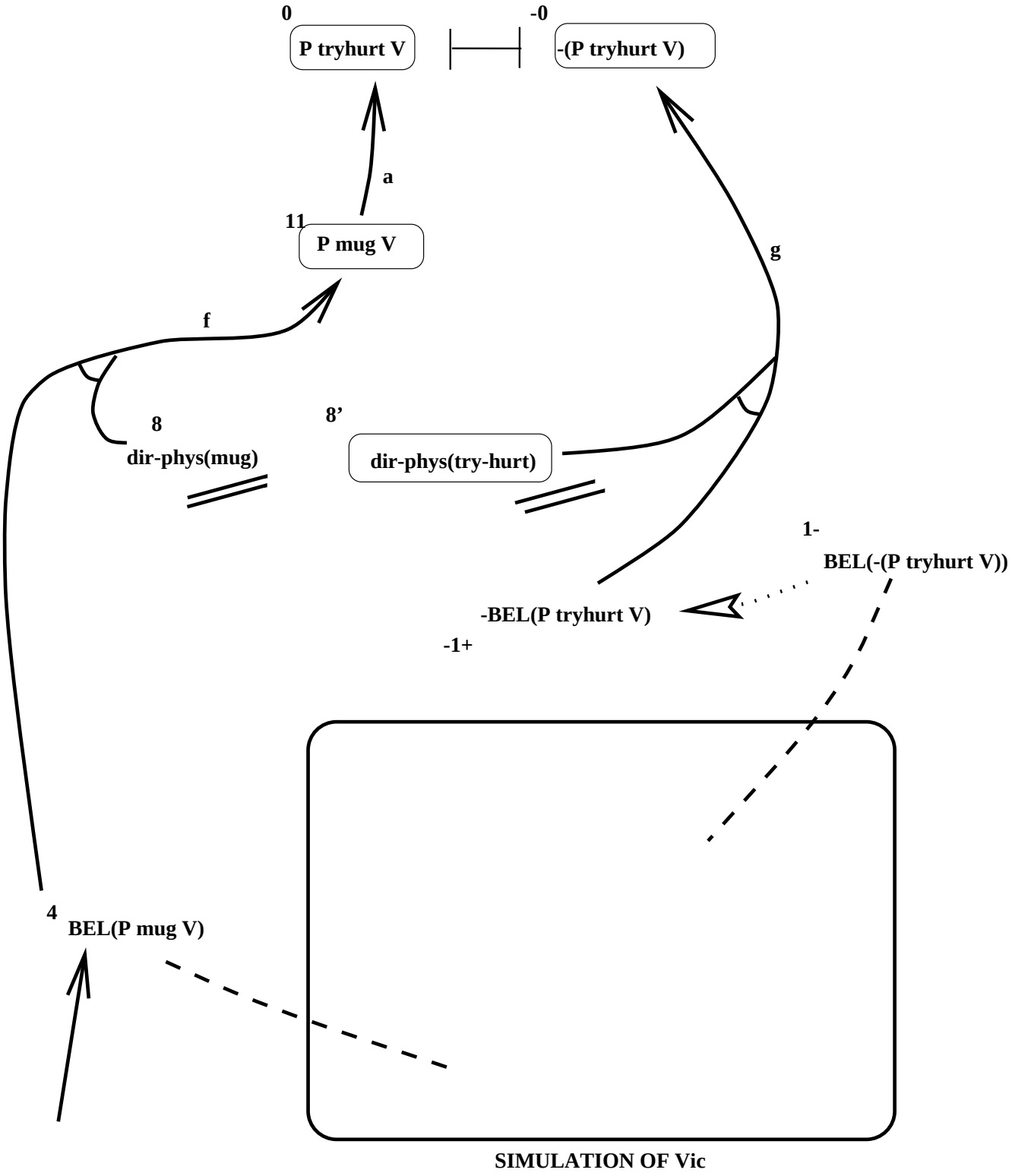

Y:R
-Y:R
-Y:-R
Y:-R
R
-R
X's SIMULATION OF AGENT Y
X's OWN REASONING LAYER


Y : R
-Y : R
-Y : -R
Y : -R
R
-R
Y : S
S
X's SIMULATION OF AGENT Y
X's OWN REASONING LAYER


Y BELIEVES:
Perp
mugs
Vic
LIFTED RULE:
Λ (Mugging-Injury)
Y BELIEVES:
Perp
hurts
Vic
Y BELIEVES:
Perp is
violent
Y makes inferences
from (a) and (b)
Perp
hurts
Vic
RULE:
Mugging-Injury
Perp
mugs
Vic
(a)
Perp is
violent
(b)
X's SIMULATION OF AGENT Y
X's OWN REASONING LAYER


1+
BEL(P tryhurt V)
1-
BEL(-(P tryhurt V))
-BEL(P tryhurt V)
-1+
-BEL(-(P tryhurt V))
-1-
Λ
2
P tryhurt V
-(P tryhurt V)
-2
4
BEL(P mug V)
a
3
P mug V
d
c
SIMULATION OF Vic
8
dir-phys(mug)
b
b
SAY(P mug V)
5
SAY(-(P tryhurt V))
7
-(V know P)
6


1+
BEL(P tryhurt V)
1-
BEL(-(P tryhurt V))
-BEL(P tryhurt V)
-1+
-BEL(-(P tryhurt V))
-1-
Λ
2
P tryhurt V
-(P tryhurt V)
-2
4
BEL(P mug V)
a
e
9
BEL(P is non-violent)
3
P mug V
10
P is non-violent
SIMULATION OF Vic
c
b
b
8
dir-phys(mug)
SAY(P mug V)
5
SAY(P is non-violent)
7'
6
-(V know P)


0
P tryhurt V
-0
-(P tryhurt V)
a
11
P mug V
g
f
8
dir-phys(mug)
8'
dir-phys(try-hurt)
1-
BEL(-(P tryhurt V))
-BEL(P tryhurt V)
-1+
4
BEL(P mug V)
SIMULATION OF Vic