Reset reproduction of article published in *Computational Intelligence*, Vol. 5, No. 1 (1989), pp. 1–11.
Reprinted July 1994.

Constructive Belief and Rational Representation

Jon Doyle

Laboratory for Computer Science, Massachusetts Institute of Technology
545 Technology Square, Cambridge, Massachusetts 02139, U.S.A.

Received July 19, 1988, revision accepted December 1, 1988

Abstract

It is commonplace in artificial intelligence to divide an agent's explicit beliefs into two parts: the beliefs explicitly represented or *manifest* in memory, and the implicitly represented or *constructive* beliefs that are repeatedly reconstructed when needed rather than memorized. Many theories of knowledge view the relation between manifest and constructive beliefs as a logical relation, with the manifest beliefs representing the constructive beliefs through a logic of belief. This view, however, limits the ability of a theory to treat incomplete or inconsistent sets of beliefs in useful ways. We argue that a more illuminating view is that belief is the result of *rational representation*. In this theory, the agent obtains its constructive beliefs by using its manifest beliefs and preferences to rationally (in the sense of decision theory) choose the most useful conclusions indicated by the manifest beliefs.

1 Introduction

Levesque (1984) introduced the notions of explicit and implicit belief as a way of overcoming some of the longstanding difficulties suffered by traditional theories of knowledge, which cannot treat agents of limited inferential abilities in satisfactory ways. In Levesque's theory, beliefs to which the agent assents readily (without prolonged thought) are called *explicit*, and the facts entailed by holding true these explicit beliefs are called *implicit* beliefs. The point of this distinction is that explicit beliefs need not be consistent, closed under entailment, or logically complete, thus permitting agents to exhibit some of the inferential incapacities observed in humans.

Many theories of knowledge in artificial intelligence offer their own candidates for what to consider as the agent's explicit and implicit beliefs. In doing so, most go further than Levesque's theory and draw an analogous distinction between beliefs explicitly and implicitly represented by an agent. The agent's actions are taken to depend on both sorts of beliefs; in Levesque's terms, both are varieties of explicit belief. The agent's explicitly represented beliefs appear as entries open to view in the agent's memory or database, and the agent's implicitly represented beliefs consist of readily computable conclusions not directly expressed but entailed by or in some way derivable from the explicitly represented beliefs. Let us call these two sorts of belief *manifest* and *constructive*, respectively. Many other names are also used for the same notions in artificial intelligence: synonyms for manifest beliefs include "assertions," "axioms," and "base beliefs," and sometime synonyms for constructive beliefs include "theorems," "derived" beliefs, and "inferable" beliefs. The distinction also appears in Fahlman's (1979) notions of "real" and "virtual" copies, and in the ensuing notions of "explicit" and "inheritable" properties in inheritance systems.

Putting these two distinctions together, theories of knowledge in artificial intelligence view an agent's set of beliefs as a function of its manifest beliefs, first determining the explicit beliefs from the manifest, and then the implicit from the explicit. In this way manifest beliefs represent both explicit and implicit beliefs, acting as a sort of "really explicit" belief.

The distinction between manifest and constructive belief is important because while most theories of ideal action require agents to hold infinitely many opinions about the world, the first limitation imposed by computational mechanisms in artificial intelligence is that individual states of the agent must be finitely describable. Distinguishing between manifest and constructive belief makes it conceivable that finite agents might nevertheless possess infinitely many opinions, since even finite sets of beliefs may represent, via entailment, infinitely many conclusions.[1] Indeed, most theories of belief developed in philosophy and decision theory do not draw or pursue distinctions between explicit and implicit or manifest and constructive belief, mainly because they lack strong motivations to presume finite describability of belief states. They instead pay more attention to the division between conscious and unconscious belief, arguably a different distinction than either explicit/implicit or manifest/constructive (but see (Konolige 1986)).

This paper considers some limitations of one element of this conception, namely the idea that the derivation of constructive and implicit beliefs from manifest beliefs is in substance logical, describable by a logic of belief. The logical conception of constructive belief makes it difficult to treat properly some common cases in which the agent's beliefs are incomplete or inconsistent, or in which the agent has limited computational resources. We suggest an alternative conception, based on the notion of *rational representation*, that overcomes these limitations in natural ways. In rational representation, the composition of the agent's constructive beliefs depends on its preferences about its states of belief as well as on its manifest beliefs. The representations explicitly possessed by the agent are not themselves viewed as functioning beliefs, but only as materials or *prima facie* beliefs from which the agent rationally constructs the beliefs on which it bases its actions. The agent's explicit beliefs are identified with its constructive beliefs, rather than with the sum of constructive and manifest beliefs, so that the set of explicit beliefs may be either more or less than the beliefs entailed logically by the manifest ones. That is, we keep the idea that manifest beliefs represent explicit and implicit beliefs, but change the nature of the representation function from logical closure under derivations to rational choice.

Once we have developed the theory of rational representation for the case of manifest and constructive belief, it is natural and straightforward (and in certain respects, almost mandatory) to broaden the theory to cover the parallel cases of manifest and constructive varieties of other mental attitudes, such as desires or goals, preferences, intentions, plans, policies, and procedures. The theory also provides natural reasons for viewing implicit beliefs as rational constructs from manifest attitudes rather than simply the beliefs logically entailed by the explicit beliefs.

Theoretically, rationality serves as an ideal every bit as attractive as logicality. Some theories of knowledge view deviations from logicality as "performance" failures that do not reflect upon the suitability of the logical "competence" theory. In contrast, the theory of rational representation views some common sorts of deviations from logicality as part of the competence theory, not as mere failures in performance. In fact, rational representation is just one element of a rich interplay between the various forms of rationality in representation and reasoning. A comprehensive treatment of these forms and their interactions is beyond the scope of this paper, but is pursued in (Doyle 1988a). The result is a theory of bounded rationality in which limited forms of rationality are themselves used to shape the overall limits on the agent's rationality.

[1] The sense in which we use "represent" here is distinct from the sense in which the agent's beliefs represent something about the agent's world. We do not mean to treat the latter here.

2 Logical and nonlogical representation

To better understand the nature of constructive beliefs, we begin by examining Konolige's (1985) theory of deductive belief. Konolige formalizes a deductive notion of explicit belief which we may recast (omitting the details) as follows:

1. A logical language $\mathcal{L}$ whose sentences represent, via an agreed interpretation, the contents of beliefs.
2. A set M of manifest beliefs, with $M \subseteq \mathcal{L}$.
3. A set $\mathcal{R}$ of sound derivation rules over $\mathcal{L}$ which determines a deducibility relation $\vdash_{\mathcal{R}}$.
4. A set C of constructive beliefs, with

$$C = \mathrm{Th}_{\mathcal{R}}(M) = \{p \in \mathcal{L} \mid M \vdash_{\mathcal{R}} p\}.$$

According to this view, the manifest beliefs M represent the conclusions C via closure under a set of sound deduction rules $\mathcal{R}$. In Konolige's theory, the rule set $\mathcal{R}$ is usually incomplete, so that not all entailed conclusions are constructive.

If one prefers to view belief semantically rather than syntactically, one may alter Konolige's theory to include

3′. A set $\mathcal{M}$ of models of $\mathcal{L}$ which determines an entailment relation $\models_{\mathcal{M}}$.

4′. A set C of constructive beliefs, with

$$C = \mathrm{Th}_{\mathcal{M}}(M) = \{p \in \mathcal{L} \mid M \models_{\mathcal{M}} p\}.$$

For example, Levesque (1984) employs special models called situations to obtain limited notions of explicit belief.

Viewed abstractly, each of these theories describes the constructive beliefs C as a function of the manifest beliefs M, that is,

$$C = f(M) \tag{1}$$

with $f = \mathrm{Th}_{\mathcal{R}}$ in the one case and $f = \mathrm{Th}_{\mathcal{M}}$ in the other.

As evidenced by Konolige's theory of deductive belief, some theories use logic as a theory of thinking by taking mental objects to be sentences in a logical language and mental operations to be inferences in a formal logical system, with the agent's constructive beliefs then some of the logical consequences of its manifest beliefs. This conception of representation is attractive since the fundamental idea underlying the notion of logical entailment or derivability is that of identifying what is implicit in given facts. But it does not follow that all interesting means of identifying constructive beliefs must be forms of logical derivations or entailment. In fact, there are strong reasons having to do with how the agent handles incomplete and inconsistent beliefs for thinking that constructive beliefs are, in some cases, both more and less than the deductive consequences of the agent's manifest beliefs, that is, that constructive beliefs can be supralogical or sublogical. Most of these reasons and the examples on which they are based are fairly well known, but have not been fully incorporated into theories of knowledge since they are not easily stated as forms of logical derivations. For example:

- Some natural categories of conclusions do not follow logically from the manifest beliefs, yet pervade commonsense reasoning. These include taxonomic defaults and circumscriptive inferences. If the theory of constructive belief is to incorporate such unsound conclusions, the derivation function cannot be purely deductive.

- Harman's (1986) "immediate implications" also make for supralogical constructive beliefs. For our purposes, immediate implications are just logically unsound inference rules, such as "If today is Tuesday, tomorrow is Wednesday." Of course, such rules might be cast as ordinary implications, but that changes the character of constructive belief. Immediate implications cannot be manipulated or combined in as many ways as can statements, so when cast as inference rules they make for weaker and less complete sets of constructive beliefs.

- Some theories of constructive belief attempt to achieve a degree of psychological accuracy by mirroring inferential limitations that humans suffer. Thus if humans do not seem to be able to draw some inference on their own, such theories of constructive belief should not prescribe the undrawn conclusion as a constructive belief. For example, many sorts of inferential limitations are thought to follow from the strategies or procedures used to conduct reasoning and the limits on the time, memory, or other resources available to these procedures. In such cases, the constructive beliefs should not be closed under Modus Ponens, since logic describes logically possible inferences, not necessarily the economically feasible inferences. Harman's immediate implications are also intended to capture such limitations on inferential capabilities, and so also yield sublogical constructive beliefs.

- Some of the systems developed in artificial intelligence provide for retracting or avoiding some manifest beliefs by making them defeasible. For example, the unsound conclusions indicated by taxonomic defaults may be overridden by more specific but contrary beliefs. Such defeated assumptions are manifest beliefs omitted from the constructive beliefs upon explicit command.

- Finally, deductive theories of constructive belief have something nontrivial to say only if the manifest beliefs are consistent, and so are unable to handle the inconsistent beliefs that arise regularly in artificial intelligence systems. When the agent detects inconsistencies in its manifest beliefs, one possible response is to first select some consistent subset upon which to reason and to then make the constructive beliefs the consequences of the consistent subset alone, thus omitting the remaining manifest beliefs from the constructive beliefs.

In each of these cases there has been considerable debate about how to define or view constructive belief, since ordinary logic provides no guidance. Some theories of belief suggest deriving constructive beliefs using a nonstandard or deviant logic instead of ordinary logic. We may easily get different theories of constructive belief by using the rules of different logics, or by using restricted classes of models rather than all possible worlds. For example, Moore (1985) employs standard epistemic modal logics; Konolige (1985) permits one to use different incomplete sets of ordinary sound rules; Levesque (1984) employs situational models connected with relevance logic; and Shoham (1987) presents a version of circumscriptive entailment based on the concept of minimal models. Except for Shoham's, these theories of belief all agree on the essentially deductive nature of constructive beliefs. There is no requirement that either manifest or constructive beliefs be complete, but the derivation rules are required to be sound (truth preserving), whether according to ordinary models or, as with Shoham's theory, according to a restricted class of models. In the latter case, the logic has embedded concepts, and may have important nonstandard characteristics (e.g., proof procedures do not always exist; see especially (Barwise 1985)).

Even though the logical conception of manifest and constructive belief is, through the use of nonstandard logics, wide enough to incorporate many interesting theories, it is not free of difficulties. The first problem is whether every interesting representation function f arising in realistic applications can be characterized in this way with suitable choices of rules $\mathcal{R}$ or models $\mathcal{M}$. If $\mathcal{M}$ must be a subset or superset of the set of ordinary models, this seems unlikely.

Second, even if one may always find some deviant logic which describes a desired f, such a logic need not be illuminating. Such a logic may precisely characterize a class of constructive beliefs, but what we really seek are theories that reflect the nature of independently justified concepts: theories that not only precisely define the conclusions of interest, but also explain why these conclusions are of interest rather than some other sorts of conclusions. Thus if our aim is to understand nonlogical representations as thoroughly as possible, we must find independent justifications of the suitability of deviant logics.

The third problem is that some theories of belief do not fit the functional mold of Equation (1) at the outset, as they associate several possible distinct sets of constructive conclusions with each individual set of manifest beliefs. One such theory is that embodied in Doyle's (1979) reason maintenance system, which has received formal treatments including McDermott and Doyle's (1980, McDermott 1982) nonmonotonic logics, Reiter's (1980) logic of defaults, Moore's (1983) autoepistemic logic, and Doyle's (1983) theories of reasoned assumptions. The problem here is not that the agent's beliefs may have different models (or, as in circumscription, minimal models) in which different things are true. That is the usual case in both ordinary and deviant logics, and is why logic defines entailment as what is true in each of the models in the given class $\mathcal{M}$. Entailment, by that definition, always yields a single set of conclusions. The problem here is instead that in some theories there are multiple, incompatible sets of conclusions, not just multiple incompatible models. If we wish to accommodate these ambiguous or nondeterministic sorts of theories in theories of constructive belief, we must generalize the representation function f to a representation relation or correspondence F, rewriting Equation (1) as the condition

$$C \in F(M). \tag{2}$$

3 Rational representation and reasoning

We suggest that a more illuminating theory of constructive belief is that it is a case of rational representation, in which the agent treats its manifest beliefs as specifications of its constructive beliefs and rationally chooses how to interpret or construe (hence "constructive") these specifications to get the constructive beliefs. Rational representation means that the agent has ways of comparing the different sets of implicitly represented beliefs it is capable of constructing, and chooses answers to queries based on the set it expects to give the best results in reasoning and representation. This does *not* mean the agent must actually construct and compare all these sets of beliefs, only that its behavior conform to this interpretation. (We postpone discussion of ways of mechanizing rational representation until Section 8.)

Although just what choosing "rationally" might mean in this context is a topic of great depth, to convey the basic idea as simply as possible we employ here the standard conception of rationality, in which a choice is rational if it is of maximal expected utility (see, for example, (Jeffrey 1983)). In addition to ordinary beliefs and desires which merely evaluate events as true or false and good or bad, the agent is presumed to possess sets of comparative attitudes, including probabilistic comparisons of events as likelier or less likely, and preferential comparisons of events as better or worse. The standard theory of rational choice takes the agent's sets of judgments of relative likelihood and desirability or utility to be consistent and complete enough that they may be represented by

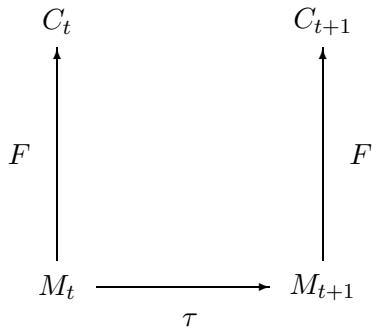

Figure 1: Representation and reasoning displayed along orthogonal dimensions: representation as a relation F between synchronic sets of manifest and constructive beliefs (M_t and C_t), and reasoning as a relation τ between diachronic sets of manifest beliefs (M_t and M_{t+1}).

numerical functions giving degrees of probability and utility for each possible event or consequence of choices. A choice is then defined to be of maximal expected utility if the total utility of the consequences of making that choice, when weighted by the probabilities of those consequences, equals or exceeds that of any alternative.

Before we proceed to justify this view of representation, note first that rational comparison of alternatives may yield several maximally good possibilities, none of which dominates the others. This means that the theory of rational representation sometimes says there are several sets of constructive beliefs corresponding to a single set of manifest beliefs. By basing its ambiguities on the familiar ambiguities of rational choice, rational representation explains in a motivated way why the relation between manifest and constructive beliefs is a correspondence F rather than a function f.

In brief, the fundamental reason the representation relation reflects rational choice rather than logical entailment is because the representation involved in constructive belief is not a passive relationship but instead an activity, one step in the more general process of reasoning in time. The usual normative theory for all activities is rational choice; hence the normative theory of constructive belief is also rational choice.

Let us examine these claims in more detail. It is natural to view representation and reasoning as relations along orthogonal dimensions when we view an agent at the knowledge level (Newell 1982). In one dimension, we view the representation relation F as an instantaneous relation between synchronic sets of manifest and constructive beliefs. In the temporal dimension, we view the state changes of the agent as a relation τ connecting each state of the agent with its possible successors. If we take the agent's states to be memory states consisting of sets of manifest beliefs, then τ is a relation between diachronic sets of manifest beliefs. More precisely, τ denotes the purely "internal" portion of the agent's actions, as we ignore effects in the agent's environment. We also assume for simplicity that the agent's steps of reasoning do not depend on its history or environment but only on its current memory state. We may display the two dimensions pictorially as in Figure 1, in which M_t and C_t denote the sets of manifest and constructive beliefs at time t.

But Figure 1 is misleading, suggesting that constructive beliefs are mere epiphenomena of the agent's activity. Instead, the purpose of constructive belief is to provide the basis for the agent's

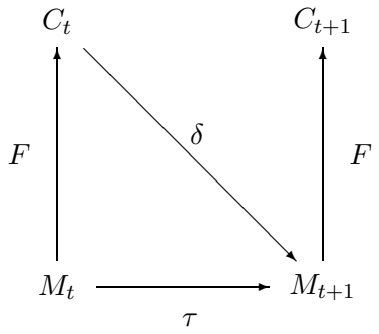

Figure 2: Constructive belief displayed as one part of a step of reasoning. Successive manifest states are calculated by δ on the basis of constructive beliefs determined by F.

actions and steps of reasoning. The mechanization of each full step of reasoning from M_t to M_{t+1} involves computing from M_t both those portions of C_t relevant to determining the possible successor states and those portions of C_t relevant to choosing the next state from among these possible successors. The relation F stands for the computation of parts of C_t, and we write δ to denote the operation of identifying and choosing possible successor states on the basis of the information in C_t. Formally, we view δ as a relation that connects sets of constructive beliefs at one instant with the possible sets of manifest beliefs at the next. (δ thus reflects the fact that constructive beliefs are changed only indirectly through changes in the manifest beliefs.) The result is that contrary to the impression given by Figure 1, τ and F are not defined independently, but instead τ "factors" through F, with the interdependence expressed by the equation

$$\tau = \delta \circ F. \tag{3}$$

That is, $M_{t+1} \in \delta(C)$ for some $C \in F(M_t)$. In particular, we have $M_{t+1} \in \delta(C_t)$. We display this more accurate picture of reasoning in Figure 2.

Ideal theories of reasoning and representation must set standards for each of these relations. The first and easier of these cases is that of reasoning. As displayed in Figure 1, reasoning is an activity in time. Reasoning or inference in this psychological sense is not to be confused with the competing logical sense of atemporal proofs, derivations, or implications within a formal logical system. The latter relations are not temporal, but are purely mathematical. As Harman (1986) puts it, inference is not implication: reasoning and inference are activities, and proofs in a logic are distinct from the activity of constructing proofs. Since the usual standard for judging activities is rationality, and since reasoning is an activity, the natural standard for reasoning is that each step of reasoning (each change of state) should be rational according to the reasoner's probabilistic expectations and preferences concerning their effects.

Conversely, it should be clear that logic has little to say about how to reason. Logical rules of inference tell us nothing about what beliefs to adopt or abandon at each step, which is what any theory of reasoning as an activity must specify. Instead, logic merely tells us what is consistent or entailed by given hypotheses. Thus logic is not, and cannot be, the standard for reasoning. Another way of putting this is that while logic may be used to *formalize* reasoning, it does not thereby *explain* reasoning. Logic, of course, may be employed to formalize psychological theories:

for example, in formulating closure and consistency properties of the agent's instantaneous sets of attitudes, or in axiomatizing the possible trajectories of the agent's states. But this use of logic is not particular to psychology. Logic may be used in the same way to formalize meteorology or any other subject matter. Using logic to formalize psychology does not thereby make mental operations inherently logical operations, any more than using logic to formalize meteorological events thereby makes thunderstorms inherently logical operations. (See also (McDermott 1987).)

Now consider the case of ideal standards for representation. There are reasonable temptations to associate the representational dimension of reasoning with the logical sense of inference. Doing so is the basis for the logicist approach to artificial intelligence (see (Genesereth and Nilsson 1986)), which makes the representation relation F purely deductive and (to the extent that it discusses rational choice at all) places all forms of rational choice into the relation τ. But it is hard to view F in this way when we consider the purpose of constructive belief in mechanized reasoners. While at the knowledge level F is an instantaneous, effortless relation, at the level of mechanization it is a computational, costly operation. Indeed, the knowledge level is an abstraction precisely intended to hide this character of F. Since F constitutes the first "half-step" of each step of reasoning and incurs computational costs and computational benefits, the natural standard for judging its constructive acts is rationality.

Of course, the dependence of reasoning on computational representation does not in itself foreclose the logicist hope of making representation purely deductive by pushing all choice into reasoning. But as a practical matter, this separation does not appear to be feasible, as typical F and δ calculations often use many of the same rules or procedures. Of course, some algorithms used in these computations may be quite distinct, with F, for example, involving marker propagation in an inheritance network, and δ involving production rules, conflict resolution, and reason maintenance. But usually the dividing line is unclear. Sections 4, 5, and 6 elaborate several apparently unavoidable ways in which representation and reasoning overlap. These sources of overlap, involving the effects of incomplete and inconsistent beliefs and limited inferential or computational resources, show that the notion of practical representation is intimately connected with the notion of rational choice; that if the standard for reasoning is that τ constitutes rational choice of successive states, then the standard for representation is that F constitutes rational choice of constructive beliefs.

4 Rational assumptions from incomplete beliefs

One consequence of employing finite representations is that the agent's knowledge of most subjects is usually incomplete. This poses a problem, for the missing information is often needed in choosing how to act. One response to this problem is skepticism, refusing to make any decisions for which the needed information is lacking. But skepticism is not always reasonable. Sometimes the more appropriate response is to make guesses or assumptions about the missing information. In such cases, we wish to use rationality as a standard for adopting assumptions by saying that an assumption should be adopted if the expected utility of holding it exceeds the expected utility of not holding it.

The classic example showing a case of skepticism to be irrational goes by the name of "Pascal's wager" and concerns religeous belief. Pascal (1662) framed his problem of belief in God in the following way: he can either believe or doubt the existence of God, and God may either exist or not exist. If God exists and Pascal believes, he gains eternal salvation, but if he doubts he suffers eternal damnation. If God does not exist, belief may lead Pascal to forgo a few possible pleasures during his life that doubt would permit him to enjoy. We may summarize these evaluations in the decision matrix shown in Figure 3, where ϵ represents the finite amount of pleasure enjoyed or

	God exists	doesn't
Believe	$+\infty$	$-\epsilon$
Doubt	$-\infty$	$+\epsilon$

Figure 3: Pascal's decision about belief in God.

forgone due to belief during Pascal's life. Of course, these same quantities modify the first column as well, but finite modifications to infinite quantities are negligible. As long as God's existence is not judged impossible, the expected utility of belief is $+\infty$, dominating the expected utility of doubt, $-\infty$. This convinced Pascal that skepticism was not a viable alternative for him.

Much later, William James (1897) made the case that skepticism is also irrational in many mundane situations. Today, his theory of the "will to believe" is one of the pillars of artificial intelligence practice, for knowledge representation and reasoning systems are filled with mechanisms for making assumptions in response to incomplete information. These mechanisms, including taxonomic defaults, threshhold probabilities, and nonmonotonic or circumscriptive proof procedures, are ordinarily not presented in terms of rational choice, and their mechanization usually involves no decision theoretic calculations (see Section 8). But when closely examined, they are clearly based on rational responses to computational problems involving incomplete information (see (Doyle 1983) and (Shoham 1987)). Taking action requires information about the available actions, about their expected consequences, and about the utility of these consequences to the agent. Ordinarily, obtaining such information requires effort, it being costly to acquire the raw data and costly to analyze the data for the information desired. The effort required for this is far too great in much of commonsense and expert reasoning, so the only feasible route for the agent is to ignore most possibilities, relying on reasonable assumptions until they prove wrong. To minimize or avoid information-gathering and inference-making costs, artificial intelligence makes heavy use of heuristics—rules of thumb, defaults, approximately correct generalizations—to guess at the required information, to guess the expected conditions and expected conclusions. These guesses are cheap, thus saving or deferring the acquisition and analysis costs. But because they are guesses, they may be wrong, and these savings must be weighed against the expected costs of making errors. Each particular case of default reasoning appearing in artificial intelligence represents a judgment that it is easier to make an informed guess and often be right than to remain agnostic and work to gather the information. These judgments are frequently made easier by recognizing that when the actions in question are mere steps of reasoning, problem-solving, or search, with no immediate external effects, it is usually the case that errors are easily correctable and ultimately inconsequential, and that the information needed to correct or verify these guesses may well become available later anyway in the course of further reasoning. In other cases, defaults are avoided, either because there is no information available to inform the guess, or because even temporary errors of judgment are considered dangerous.

Of course, not all possible assumptions are rational, and not all mechanisms for making assumptions always yield rational assumptions. For example, the tool logic offers for treating incomplete information is reasoning by cases, and in some situations it is possible to view assumption-making as ordered exploration of each of the logically possible cases. But proper evaluation of possible

beliefs means taking both utility and probability into account. This means it can be rational to assume and explore one case but not to assume or explore any of the remaining cases if the first one proves unhelpful. One might be informed that a book one seeks is stocked by one of three stores A, B, and C, with the chances being 2/3, 1/6 and 1/6 respectively. Other things being equal, it is rational to visit the stores in either the order ABC or ACB. But if B and C are located in Beirut or the South Bronx, it is for most people rational to give up the exploration if A proves fruitless.

Similarly, it is not necessarily rational to base assumptions purely on utilities, assuming something as long as its utility exceeds some threshold, regardless of the probability of its being true. This is called wishful thinking, and is deservedly avoided. It is also not necessarily rational to draw conclusions just as long as their probabilities exceed some threshold value, or if they hold in the limiting case of small uncertainties (as in (Pearl 1988)). For example, all tautologies have maximal probability, but most are worthless to the reasoner. Except in carefully designed applications, belief based purely on degree of likelihood is usually a mistake. This mistake has no notorious name, and perhaps not coincidentally, has enjoyed some popularity in artificial intelligence.

The most striking result in this regard is that when both probability and utility are taken into account, it is in some cases rational to make assumptions expected to be false.[2] For example, if one is cast adrift in the middle of the Pacific Ocean, one may expect that there is only a remote chance of being rescued. But one may also know that people who despair die quickly, while people with an unconquerable belief in their salvation can survive long periods of extreme privation. Since the longer one survives the greater the chance of rescue, belief in rescue often turns out to be self-fulfilling. Thus while the probability of rescue is low, the utility of belief in rescue is high, so it is rational to believe each day (rejecting all doubts) that today is the day of rescue.

Of course, Pascal's decision and other cases of adopting beliefs are usually viewed as steps of reasoning, as temporal changes in manifest beliefs. But most mechanisms for making assumptions explored in artificial intelligence involve rules for making classes of assumptions rather than individual assumptions. We may assimilate the two cases theoretically by evaluating individual default rules as individual assumptions. Default rules can be used in temporal steps of reasoning to adopt individual assumptions, but it is even more common to use them in calculating constructive beliefs. The taxonomic defaults employed in inheritance systems offer perhaps the clearest example of constructive uses of rules of assumption.

The strong practical motivations for using default rules in defining and computing constructive beliefs defeat attempts to make the representation relation F purely deductive. Constructive belief is most useful when it includes the commonly needed assumptions that could be readily made in proper steps of reasoning, and the proper assumptions to make are those whose expected benefit is greatest.

5 Rational interpretations of inconsistent beliefs

If agents ordinarily have to act on the basis of incomplete information, at other times they have to act on the basis of inconsistent information. Conflicting information may reach the agent through many different sensors and informants. In the simplest case, inconsistencies appear because the agent's beliefs are drawn from several experts who disagree about the facts (in which case the inconsistency may be manifest), or who think they agree because the inconsistencies in their views are too subtle to detect (in which case the inconsistency may not be explicit). It is not always possible to decide immediately which information is correct and which is false and accept the true

[2]More generally, one may judge lying rational just as one may judge honesty rational. Certainly lying to oneself would not be as common as it is if it did not offer some sort of large reward.

and reject the false. Instead, many conflicts require extended reasoning or investigation to resolve. Meanwhile, the agent must continue to decide what to do in spite of these conflicts.

Logic does not appear to offer much guidance about what to do when the manifest beliefs are inconsistent, since every conclusion is entailed by an inconsistent set of hypotheses. (Indeed, manifestly inconsistent belief constitutes the one case in which an agent can always immediately answer every question put to it.) At best, logic says we should confine attention to a consistent set of beliefs and act on that basis. Ordinary logic does not say which consistent sets to consider, but some special logics have been developed specifically for reasoning with inconsistent hypotheses. Some draw only those conclusions holding in every maximal consistent subset of the hypotheses (see, for example, (Rescher 1964)). Others, such as relevance logic, draw all conclusions entailed by any consistent subset and tag each conclusion with the subset supporting it (see, for instance, (de Kleer 1986) and (Martins and Shapiro 1983)).

Unfortunately, none of these purely logical approaches really offer any guidance in acting with inconsistent beliefs. Acting on the basis of a consistent subset of the manifest beliefs is reasonable, but none of the logics help decide which subset to use. Fortunately, however, there are many similarities between the cases of acting on incomplete information and acting on inconsistent information. With inconsistent beliefs, the agent faces a situation of ambiguity, just as in the case of incomplete beliefs, but now the alternatives are the consistent subsets of beliefs rather than the disjuncts of some disjunction. Moreover, just as in the case of incomplete beliefs, there is often information available about which choice is better, about which subset to prefer. For example, if neighborhood gossip is found to be in conflict with Newtonian mechanics, few would reject Newton and accept the gossip. The major difference between the cases of incomplete and inconsistent belief is that while the agent may rationally refuse to make any assumptions in some cases of incomplete belief, an agent possessing conflicting beliefs cannot refuse to choose a basis for action.

A choice of basis for action can be approached in the same way as a choice of assumptions, as an instance of rational representation. In this approach, the agent applies its preferences about consistent subsets of its manifest beliefs to rationally choose one, and then uses this subset to choose a consistent body of constructive beliefs (though these decisions need not be separate decisions since the subset may be chosen so as to yield a desired conclusion). Rational representation allows the possibility that the agent may select different subsets of the manifest beliefs for different actions, even if the manifest beliefs remain constant. In such cases, the logical inconsistency of the agent's beliefs is exhibited in the "inconsistency" of the agent's actions in unchanging circumstances. This "inconsistency in action" can be minimized by employing the notion of conservative revisions of constructive beliefs, as discussed in Section 8.

In choosing consistent sets of constructive beliefs based on inconsistent manifest beliefs, we may think of the constructed beliefs as "representing" the inconsistent manifest beliefs for the purpose of the action at hand, just as we think of the elected officers of a political organization as representing its membership. This sense of "representation" is different from the one we have been studying so far. In the first sense, the manifest beliefs represent the constructive beliefs: in the second, the constructive beliefs also represent the manifest beliefs. This second sense is not merely a curious coincidence of words. The most common and unavoidable sort of inconsistent knowledge employed in artificial intelligence is not inconsistent belief, but conflicting default rules and the conflicting preferences they embody. We cannot pursue the point here, but reasoning with conflicting default rules turns out to be formally the same problem as decision-making or self-government by groups of people, thus justifying the second sense of the term "representation." See (Doyle and Wellman 1988) for more on this connection between rational knowledge representation and theories of social or political decision-making.

6 Rational expenditure of limited resources

If an agent's manifest beliefs and derivation rules are logically consistent and complete, then the consequences of these beliefs are decidable. But even when this is the case, it may not be feasible to consider all these consequences as constructive beliefs if only limited resources of time and memory are available to compute them.

Two major approaches have been explored toward defining constructive beliefs in situations of limited resources. One approach is exemplified by Konolige's (1985) formalization of deductive belief, as discussed in Section 2. The basis of Konolige's theory is in describing constructive beliefs as the closure of the manifest beliefs under an incomplete set of deductive rules, where the rule-sets so employed are restricted to those that can be mechanized within the prescribed resource bounds. The second approach is to use logics of belief which explicitly incorporate descriptions of resources into the description of states. Instead of describing what beliefs follow from others, these logics describe which constructive beliefs follow from the manifest beliefs and given quantities of resources, for example, in terms of how many applications of Modus Ponens are needed to derive a conclusion. There is some overlap in these two approaches, as in some cases it may be possible to find an incomplete set of rules that exactly captures the effect of a specific sort of resource limitation.

While theories of belief that take limitations on reasoning abilities into account are a step in the right direction, theories that describe these limitations purely in terms of resource bounds suffer from serious difficulties. The first difficulty is that the quantities of resources available to the agent need not be well defined. The agent's stock of resources may change through consumption or through changes in the agent's environment. In addition, some resources may be augmented as well as consumed by the agent's actions. Indeed, the supplies of the most important mental resources are not fixed, but are instead what the agent makes them through investment of effort in their improvement or destruction. For example, deadlines can sometimes be postponed to gain more time, and effective memory capability can be increased by reorganization or culling of memory, or by augmentation with external memory aids. Such fluidity in the resources available makes inapplicable logics of limited belief such as Konolige's or Davis' (1981) logic of obvious inferences, since these logics reflect fixed limits to reasoning. There is no general and natural yet static logic of limited belief.

The second, and more telling difficulty is that the agent may have the license and resources to draw a conclusion, but no interest in (or even a definite antipathy toward) drawing it. Such undrawn conclusions are not simply a matter of competence and performance, for we would think an agent incompetent if it could not avoid things it intends to avoid and has the power to avoid. Yet the very idea of avoided conclusions contradicts the implicit assumption of many theories of resource-limited belief that all conclusions that can be drawn within the available resources should be drawn, that having more beliefs is always better.

Defeasible reasoning, which underlies many common forms of reasoning and representation, provides many examples of conclusions avoided because other information indicates the reasons for deriving them to be inapplicable. But the more striking instances of avoided conclusions involve deliberate ignorance. Just as it is not always rational to be skeptical, it is not always rational to be credulous, and deliberate ignorance is called for when skepticism is rational. Deliberate ignorance is foreign to scientists, who are trained to want to know everything. Indeed, Good (1967) has proven that it is not rational to ignore available information when the costs of collecting and using it are negligible. But these costs are often *not* negligible. For example, deliberate ignorance is common in the everyday lives of people who often would rather not know something they could easily find out, since some sorts of information can make the learner very unhappy, in which case learning carries a

	Taxes overlooked	none overlooked
Read	$-\infty$	$-\epsilon$
Don't	$-\$$	0

Figure 4: Rascal's decision about reading the tax code.

great emotional cost. This sort of behavior is not merely a human foible. In computational systems with autonomous subsystems, and in bureaucracies with strictly enforced procedures, a piece of information relevant to the agent's aims might be easy to acquire, but nevertheless be rationally avoided because its acquisition might automatically trigger costly processing (e.g., endless forms, expensive investigations) the agent wishes to avoid.

We may modify Pascal's wager slightly to provide a concrete example of rational ignorance. Suppose Pascal's cousin Rascal is involved in shady financial dealings, and that while he believes that there may be tax codes that call for extra taxes on his earnings, he would (being an inveterate cheater) sometimes avoid paying them even if he knew about them. Rascal also knows that in his country the penalty for deliberately evading specific taxes is death, while the penalty for inadvertently overlooking esoteric taxes is merely monetary. One day Rascal comes upon the ten volumes of the tax code in the library. He then faces the question of whether to read the tax code to see if he is overlooking some taxes. Figure 4 summarizes Rascal's evaluation of the alternatives of reading or not reading the tax code, under the assumption that his shady dealings will someday lead to his taxes being audited. We let $-\infty$ represent the disutility of death, $-\$$ represent the disutility of the taxes due plus the fine for overlooking taxes, and $-\epsilon$ represent the disutility of spending the time to read the ten volumes of tax code. Rascal rationally concludes that if there is any chance of his overlooking any taxes that he would not pay, he is better off not knowing about them. (For a game-theoretic treatment of less contrived examples of rational ignorance, see (Cave 1983).)

We may summarize this discussion in the observation that resource-limited reasoning is really a code-word for the economics of reasoning, for the rational allocation of mental resources. The agent's ability to come to specific conclusions, as well as its probability of coming to these conclusions, depends on the agent's preferences as well as its beliefs, as these preferences determine or influence both the types and amounts of resources available to the agent, and the interest or motivation of the agent toward making specific inferences. Applied to the theory of constructive belief, this observation means that extant approaches to constructive belief are unnecessarily rigid and possibly wasteful, as they focus on the special case of an agent bound to construct every belief it can within the limits of its resources. The theory of rational representation offers a more general and natural approach, in which there may be several sets of constructive beliefs corresponding to a single set of manifest beliefs. Even when fixed amounts of resources are available for computing constructive beliefs, the representation service may allocate these resources in several different but equally rational ways, each way yielding a different set of constructive beliefs.

7 Constructive attitudes and implicit belief

The preceding has made a case for thinking of constructive belief in terms of rational representation. It is not hard to see that a similar view is reasonable for other attitudes involved in the agent's knowledge. That is, the agent's manifest beliefs, plans and preferences are just the starting point for calculation of its constructive beliefs, constructive plans, and constructive preferences. The same considerations motivating finite representations of sets of beliefs motivate finite representations of large sets of preferences and plans or policies.

Having all types of attitudes represented in manifest and constructive forms introduces subtle problems into the very definition of rational choice of constructive attitudes. If preferences may be constructive, rather than always manifest, then the proper definition of constructive attitudes should involve rational choice with respect to the constructed preferences themselves, not just with respect to the manifest preferences. This makes each choice of constructive beliefs what Jeffrey (Jeffrey 1983) calls a *ratified* decision, that is, a choice that is rational with respect to the set of attitudes resulting from that choice. For example, if sets C and C' are both rational constructions according to the preferences in M, then M does not prefer C over C' or vice versa. But if C contains a constructive preference for C' over C, then only C' could be a ratified choice from M, since C says of itself that it is not the best choice.

This same line of reasoning suggests reconsidering the theoretical relation between explicit and implicit belief. Recall that Levesque (1984) defined implicit beliefs to be just those facts entailed by the agent's explicit beliefs. This definition accords well with the usual meaning of the term "implicit," and provides a natural extension of the notion of explicit belief as a partial deductive closure of manifest axioms. In addition, this definition of implicit belief is motivated by a view of reasoning in which the point of "problematic" or "nontrivial" reasoning is to make explicit some of the beliefs that are presently only implicit. A central feature of such reasoning, according to this view, is that it leaves the set of implicit beliefs invariant, changing only the set of explicit beliefs. But it is not clear that we can retain this view of implicit belief in the setting of rational representation.

First consider what happens when constructive attitudes are made manifest. The constructive preferences may contain preferences that are not manifest. Making one of these manifest may have the effect that some previously rational sets of constructive belief are no longer rational, as they did not contain the added preference and are ruled out by it. However, the added preference cannot rule out the set of constructive attitudes from which it was drawn, since that set was a ratified decision. If we assume that the added preference does not allow the construction of any new attitudes, then this addition leaves invariant the set of constructive attitudes.

Now consider the case of implicit attitudes. Just as the constructive preferences may contain preferences that are not manifest, the constructive preferences may entail some implicit preferences that are not manifest or constructive. Making an implicit preference manifest or constructive can change the ratified sets of constructive attitudes, possibly removing from consideration the very set of constructive attitudes that entailed the added preference. Thus if we want a set of implicit attitudes to remain invariant when implicit attitudes are made explicit, then we must define the implicit attitudes to be sets of attitudes that both are closed under entailment and are ratified rational choices from the manifest attitudes. This makes the notion of implicit belief more like the stable (i.e., ratified) extensions in Moore's (1983) autoepistemic logic than the consequential closure of the explicit beliefs. But it also opens the possibility of discrepancies arising between the sets of implicit and constructive attitudes, as each represents a ratified choice with respect to different sets of preferences. In particular, it may happen that the manifest attitudes give rise to several sets of constructive and implicit attitudes, but that no set of constructive attitudes is a

subset of any of the sets of implicit attitudes.

8 Mechanizing rational representation

Even if rational representation provides a more appropriate standard than deductive representation for constructive belief, as a practical matter it poses problems at least as difficult as those faced in mechanizing deductive representation. It is too hard, in general, to explicitly compare all possible sets of constructive beliefs, for the sets may be infinite in both size and number. One may not then be able to rationally choose constructive beliefs simply by retrieving expectations and preferences and calculating which state has maximal expected utility. This does not mean that the idea of rationality is useless in mechanizing constructive belief, for its role as an ideal theory is important even if it cannot be fully attained. But much work remains to be done on mechanizing rational representation, as only a few approaches have be explored, and none of these very thoroughly.

The approaches taken so far towards mechanizing rational representation attempt to approximate this ideal by using simple sorts of rational attitudes in ways compatible with their ideal meanings. Three approximation methods have been examined. The first is to consider only some of the agent's preferences and expectations at a time, and to use initially simple but progressively more complex decisions to select constructive beliefs. That is, the agent devotes a fraction of its effort to some simplified decisions about how to rationally expend the remainder, making this allocation by means of another rational choice involving restricted sorts of preferences, such as those expressed in default rules, or those employed by Smith (1985) and Russell and Wefald (1988). If need be, this simplification may be employed more than once to reach a decision. See (Doyle 1980, 1988a) for treatments of patterns of successively reflective preferences about preferences and reasoning about reasoning as elements of a theory of rationally bounded rationality.

Two related but in many ways simpler approaches towards mechanizing rational representation may be found in reason maintenance systems (Doyle 1979). The first of these is to approximate rational choice by finding some more easily computable condition that entails rationality. Reason maintenance does this by basing constructive beliefs on reasons or justifications, and requiring that constructive beliefs be grounded or possess well-founded support in terms of these reasons. Grounded sets of constructive beliefs are relatively easy to compute. When one interprets reasons as special sorts of preferences about sets of beliefs (or as comparisons of their expected utilities), it is possible to prove that grounded sets of constructive beliefs constitute rational choices in a limited sense. Specifically, grounded sets are Pareto-optimal choices with respect to the preferences expressed by reasons (Doyle 1983, 1985). But this approximation to rationality is not perfect, since not all of the agent's preferences are taken into account.

The second way reason maintenance attempts to simplify the computation of constructive beliefs is through conservative updates, where effort is saved (one hopes) by carrying over many beliefs without recomputation. Doyle's (1979) reason maintenance system did not compute constructive beliefs in precisely the sense we have been using the term, since every reason and belief it represented was manifest in memory. A more accurate view is that in that system the manifest beliefs M_t are divided into two subsets: *basic* beliefs (and reasons) B_t, and *supported* beliefs S_t, with $M_t = B_t \cup S_t$. The basic beliefs play the role we earlier assigned to manifest beliefs, in that all explicit changes to beliefs are made by changing the set of basic beliefs. The supported beliefs play the role we earlier assigned to constructive beliefs. In essence, the supported beliefs are just some of the constructive beliefs made manifest. The motivation for doing this is to ease computation of modified sets of constructive beliefs by making incremental changes in the current set of supported beliefs instead of requiring complete reconsideration of all possible sets of constructive belief. The hope here (which

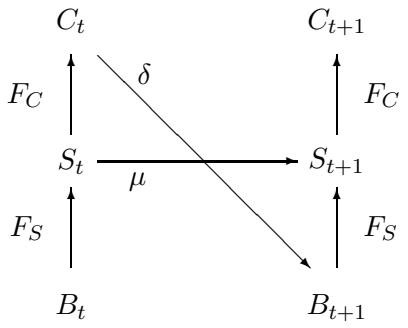

Figure 5: Reason maintenance makes manifest, as supported beliefs S_t, some of the beliefs constructed from basic beliefs B_t, with $M_t = B_t \cup S_t$. The representation relation F is split into two parts F_S and F_C. The reasoning relation δ changes only basic beliefs, and the supported beliefs are conservatively updated according to a minimal-revision relation μ.

has not yet been justified theoretically) is that such incremental revisions are in practice less costly to effect than continual recomputations of individual constructive beliefs, especially since rational choice of consequences tends to eliminate the modularity or independence of conclusions exhibited in monotonic logics. As with constructive beliefs generally, the decision about which beliefs appear manifestly as supported beliefs and which are left as purely constructive beliefs is an economic one, dependent on the role the particular beliefs are expected to play in belief construction and reasoning.

We may depict the reason maintenance approximation to rational representation as in Figure 5. Here we view the overall representation relation F as the result of two simpler relations F_S and F_C, with F_S relating basic beliefs B_t to supported beliefs S_t, and F_C relating supported beliefs S_t to constructive beliefs C_t. Reasoning changes most beliefs indirectly, through changes in the basic beliefs, and as before δ denotes the relation between constructive beliefs at one instant and changed (basic) beliefs at the next. In addition, the supported beliefs S_{t+1} are carried over from the previous state, with minimal changes made in order to maintain the relation F_S with the updated set B_{t+1}. The choice of S_{t+1} to be as "close" as possible to S_t among the alternatives in $F_S(B_{t+1})$ is denoted by μ. (We will not treat here the interesting topic of what constitutes minimal or conservative revisions, but rationality plays an important role in that concept too. See (Doyle 1988a) and (Gärdenfors 1988).) Summing up, we have $S_t \in F_S(B_t)$, $C_t \in F_C(S_t)$, $B_{t+1} \in \delta(C_t)$, and $S_{t+1} \in \mu(S_t, F_S(B_{t+1}))$.

9 Conclusion

To summarize, implicit beliefs are represented by constructive beliefs, and constructive beliefs are represented by manifest beliefs. Numerous examples were presented to show that constructive belief is best viewed as decision-theoretic notion, with constructive attitudes generally defined to be attitudes chosen so as to maximize their expected utility to the agent. The representation relation between manifest and constructive attitudes is thus rational, not logical, and the limits to knowledge are primarily economic, not logical. In particular, the agent's constructive and implicit

beliefs depend upon its preferences as well as its manifest beliefs. Since these preferences may change over time, no static logic of belief (or even of belief and resources) can capture all notions of constructive and implicit belief of demonstrated practical value or which conform to commonsense ascriptions of belief.

The involvement of choice in computational representation makes the situation in computational reasoning considerably different than the traditional logical conception of reasoning in philosophy. Philosophers have long distinguished the notions of theoretical reasoning (i.e., logic), which focuses on questions of truth, and practical reasoning (i.e., deliberation), which focuses on questions of value. Though philosophers made no distinction between manifest and constructive belief, artificial intelligence theorists have often been tempted to identify these two dimensions of reasoning with the synchronic and diachronic dimensions of Figure 1, with constructive beliefs reached by theoretical reasoning and the next state reached by practical reasoning. But this identification is not appropriate, for the distinction between manifest and constructive belief is not that between theoretical and practical reasoning. Instead, in computational reasoning both manifest and constructive beliefs exist to serve the needs of practical reasoning. Theoretical reasoning, if it exists at all in automated reasoners, is something else entirely, something not necessarily involved in the activity of reasoning.

The nonlogical nature of constructive belief is less surprising when one considers representation to be just one aspect of the organization of reasoning, for one can see immediately that logic omits the idea that reasoning has a purpose. The purpose of reasoning is not just to draw further conclusions or answer posed questions. To paraphrase Hamming, the purpose or aim of thinking is to increase insight or understanding, to improve one's view (as Harman puts it), so that, for instance, answering the questions of interest is easy, not difficult. This conception of reasoning is very different from incremental deduction of implications. Instead of simply seeking *more* conclusions, rationally guided reasoning constantly seeks *better* ways of thinking, deciding, and acting. Rational reasoning does not preserve truth but instead destroys and abandons old ways of thought to make possible invention and adoption of more productive ways of thought. Correspondingly, the purpose of representation is to offer the best conclusions to draw rather than all the logically possible conclusions, to guide the reasoner toward the useful conclusions, whether sound or unsound, and away from useless ones, whether true or false.

The task of artificial intelligence would certainly be easier were it practical to organize representation and reasoning so as to avoid non-logical assumptions and revisions, to instead involve only cumulative, logical deduction of the consequences of initial beliefs and descriptions of passing experiences. But that does not appear to be any more feasible for machines than for humans. Humans, at least, find guessing necessary because of our small, frail mental abilities. Denied complete and certain knowledge we assume our way through life, only dimly and occasionally aware through our meager senses of any reality, and even then loath to part with our cherished beliefs. We must revise and reinterpret our beliefs even if guesses are never wrong because progress in reasoning, like maturity and progress in life, requires escape from the shackles of the past. Agents whose beliefs are cumulative are unwilling to give up the past, and are condemned to repeat it endlessly. Put most starkly, reasoning aims at increasing our understanding; rules of logic the exact opposite.

Acknowledgments

An abbreviated ancestor of this paper (Doyle 1988b) appeared under the title *Knowledge, Representation, and Rational Self-Government* at the Second Conference on Theoretical Aspects of Reasoning about Knowledge, Monterrey, California, in March 1988. The author is especially indebted to Hector Levesque, Ronald Loui, and Robert Moore for their commentaries ((Levesque 1988), (Loui

1988), and (Moore 1988)) of that paper, and to Jonathan Cave, Allen Newell, Jonathan Pollock, Joseph Schatz, Richmond Thomason, Michael Wellman, and the referees for valuable comments, ideas, and suggestions.

This research was supported in part by the National Library of Medicine under National Institutes of Health Grant R01 LM04493, and in part (while the author was with Carnegie Mellon University) by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 20, monitored by the Air Force Avionics Laboratory under Contract F33615-87-C-1499.

References

Barwise, J., 1985. Model-theoretic logics: background and aims, *Model-Theoretic Logics* (J. Barwise and S. Feferman, eds.), New York: Springer-Verlag, 3-23.

Cave, J. A. K., 1983. Learning to agree, *Economics Letters*, Vol 12, 147-152.

Davis, M., 1981. Obvious logical inferences, *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 530-531.

de Kleer, J., 1986. An assumption-based TMS, *Artificial Intelligence* **28**, 127-162.

Doyle, J., 1979. A truth maintenance system, *Artificial Intelligence 12(3)*, 231-272.

Doyle, J., 1980. A model for deliberation, action, and introspection, Cambridge: MIT Artificial Intelligence Laboratory, TR-581.

Doyle, J., 1983. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, report 83-125.

Doyle, J., 1985. Reasoned assumptions and Pareto optimality, *Ninth International Joint Conference on Artificial Intelligence*, 87-90.

Doyle, J., 1988a. Artificial intelligence and rational self-government, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-124

Doyle, J., 1988b. Knowledge, representation, and rational self-government, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 345-354.

Doyle, J., and Wellman, M. P., 1988. Impediments to universal preference-based default theories, *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, San Mateo, CA: Morgan Kaufmann.

Fahlman, S. E., 1979. *NETL: A System for Representing and Using Real World Knowledge*, Cambridge: MIT Press.

Gärdenfors, P., 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, Cambridge: MIT Press.

Genesereth, M. R., and Nilsson, N. J., 1987. *Logical Foundations of Artificial Intelligence*, Los Altos: Morgan Kaufmann.

Good, I. J., 1967. On the principle of total evidence, *British Journal for the Philosophy of Science* **18**, 319-321.

Harman, G., 1986. *Change of View: Principles of Reasoning*, Cambridge: MIT Press.

James, W., 1897. *The Will to Believe and other essays in popular philosophy*, New York: Longmans, Green and Co.

Jeffrey, R. C., 1983. *The Logic of Decision*, second edition, Chicago: University of Chicago Press.

Konolige, K., 1985. Belief and incompleteness, *Formal Theories of the Commonsense World* (J. R. Hobbs and R. C. Moore, eds.), Norwood: Ablex, 359-403.

Konolige, K., 1986. What awareness isn't: a sentential view of implicit and explicit belief, *Proc. Conf. Theoretical Aspects of Reasoning about Knowledge* (J. Y. Halpern, ed.), Los Altos: Morgan Kaufmann, 241-250.

Levesque, H. J., 1984. A logic of implicit and explicit belief, *Proceedings of the Fourth National Conference on Artificial Intelligence*, 198-202.

Levesque, H. J., 1988. Comments on "Knowledge, representation, and rational self-government," *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 361-362.

Loui, R. P., 1988. The curse of Frege, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 355-359.

Martins, J. P., and Shapiro, S. C., 1983. Reasoning in multiple belief spaces, *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 370-373.

McDermott, D., 1982. Nonmonotonic logic II: nonmonotonic modal theories, *Journal of the Association for Computing Machinery* **29**, 33-57.

McDermott, D., 1987. Critique of pure reason, *Computational Intelligence*, Vol. 3, No. 3, 151-160.

McDermott, D., and Doyle, J., 1980. Non-monotonic logic—I, *Artificial Intelligence* **13**, 41-72.

Moore, R. C., 1983. Semantical considerations on nonmonotonic logic, *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 272-279.

Moore, R. C., 1985. A formal theory of knowledge and action, *Formal Theories of the Commonsense World* (J. R. Hobbs and R. C. Moore, eds.), Norwood: Ablex, 319-358.

Moore, R. C., 1988. Is it rational to be logical?, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 363.

Newell, A., 1982. The knowledge level, *Artificial Intelligence* **18**, 87-127.

Pascal, B., 1662. *Pensées sur la religion et sur quelques autres sujets* (tr. M Turnell), London: Harvill, 1962.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo: Morgan Kaufmann.

Reiter, R., 1980. A logic for default reasoning, *Artifical Intelligence* **13**, 81-132.

Rescher, N., 1964. *Hypothetical Reasoning*, Amsterdam: North Holland.

Russell, S., and Wefald, E., 1988. Decision-theoretic control of search: general theory and an application to game playing. Technical Report 88/435, Department of Computer Science, University of California, Berkeley.

Shoham, Y., 1987. Nonmonotonic logics: meaning and utility, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 388-393.

Smith, D. E., 1985. Controlling inference, Stanford: Department of Computer Science, Stanford University, Ph.D. thesis.