Implicit Knowledge and Rational Representation

Jon Doyle
April 1988
CMU-CS-88-134 (2)

Abstract: It is commonplace in artificial intelligence to draw a distinction between the explicit knowledge appearing in an agent's memory and the implicit knowledge it represents. Many AI theories of knowledge assume this representation relation is logical, that is, that implicit knowledge is derived from explicit knowledge via a logic. Such theories, however, are limited in their ability to treat incomplete or inconsistent knowledge in useful ways. We suggest that a more illuminating theory of implicit knowledge is that it is the result of rational representation, in which the agent rationally (in the sense of decision theory) chooses interpretations of its explicit knowledge.

This research was supported 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. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

1 Introduction

Though they may disagree on other points, many theories of knowledge in artificial intelligence draw a distinction between the knowledge explicitly and implicitly possessed by an agent. According to the shared view, the agent's actions depend on both its explicit and implicit knowledge, where the agent's explicit knowledge appears as entries in the agent's memory or database, and the agent's implicit knowledge consists of conclusions entailed by or derivable from the explicit knowledge, with the derivation described by a logic of knowledge or belief. This distinction is fundamentally one about representation, for it allows the agent to use a finite body of explicit knowledge to represent an infinite body of implicit knowledge.

This paper considers some limitations of one element of this conception, namely the idea that the derivation of implicit from explicit knowledge is in substance logical. We suggest an alternative conception of implicit knowledge, based on the notion of rational representation, that overcomes these limitations in natural ways. In rational representation, the implicit beliefs, for example, depend on the agent's preferences about its states of belief and on its beliefs about its states of belief as well as on the explicit beliefs themselves. The explicit representations possessed by the agent are not themselves viewed as knowledge, but only as materials or *prima facie* knowledge from which the agent rationally constructs the bases of its actions, so that its actual knowledge, as a set of attitudes, may be either more or less than the attitudes entailed logically by the explicit ones. That is, we keep the idea that the explicit knowledge represents the implicit knowledge, but change the nature of the representation function from logical closure under derivations to rational choice. In this theory, rationality serves as an ideal every bit as attractive as logicality, and moreover, provides satisfying treatments of many approaches toward reasoning with incomplete and inconsistent knowledge. In particular, in the conventional view, deviations from logicality are "performance" failures that do not reflect upon the suitability of the logical "competence" theory. In contrast, in the theory of rational inference, the common sorts of deviations from logicality are part of the competence theory, not mere failures in performance.

2 Implicit knowledge and representation

Most theories of knowledge developed in philosophy and economics do not draw the distinction between explicit and implicit knowledge, or do not make much of it if they do. (More attention is paid to the separate division between conscious and unconscious knowledge.) The distinction is important in artificial intelligence because the first limitation imposed by computational mechanisms is that individual states of the agent be finitely describable. Most theories of ideal action require agents to hold infinitely many opinions about the world, however, and distinguishing between explicit and implicit knowledge makes it conceivable that finite agents might nevertheless possess infinitely many opinions, since even finite sets of axioms may represent, via entailment, infinitely many conclusions. (This sense of representation is in addition to the sense in which the agent's knowledge represents something about the agent's world.)

We may symbolize this idea with the suggestive equation

$$I = f(E),$$

where I stands for the agent's implicit knowledge, E for the agent's explicit knowledge, and f for the function describing how explicit knowledge determines or represents implicit knowledge.

The distinction between explicit and implicit knowledge goes by other names as well in artificial intelligence. These include the distinction between assertions or axioms and theorems or derived or inferable conclusions, Fahlman's [1979] distinction between "real" and "virtual" copies, and the latter's reflection throughout work on inheritance systems as the distinction between explicit and "inheritable" properties.

3 Logical and nonlogical representation

The term "knowledge" is commonly used in both a broad and a narrow sense in artificial intelligence and related fields. The narrow sense treats knowledge as something like true belief, following a long tradition in philosophy (see [Moore 1985]). The broad sense treats knowledge as including preferential and intentional information as well as purely facutal information, thus counting information about the agent's values, plans, and procedures as part of its knowledge. We

will start by examining knowledge in the narrow sense, as it best introduces the problems that arise from the logical view of implicit knowledge, but then move to examine knowledge in the broad sense, as it best illuminates the solutions to these problems.

In the standard view of implicit knowledge, logic serves as a theory of thinking in that mental objects are taken to be sentences in a logical language and mental operations are taken to be inferences in a formal logical system, so that the agent's implicit beliefs are just the logical consequences of its explicit beliefs. Konolige [1985], for instance, formalizes explicit and implicit belief in terms of the following elements (omitting the details):

1. A logical language $\mathcal{L}$ whose sentences represent, via an agreed interpretation, the contents of beliefs.

2. A set B of base beliefs, with $B \subseteq \mathcal{L}$.

3. A set R of sound derivation rules over $\mathcal{L}$ which determines a deducibility relation $\vdash_R$.

4. A set C of derived beliefs, with

$$C = \mathrm{Th}_R(B) = \{p \in \mathcal{L} \mid B \vdash_R p\}.$$

According to this view, the base beliefs B represent the conclusions C via closure under a set of sound deduction rules R. Alternatively, one may view implicit conclusions semantically, in which case one has instead

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

4'. A set C of derived beliefs, with

$$C = \mathrm{Th}_M(B) = \{p \in \mathcal{L} \mid B \models_M p\}.$$

Each of these theories clearly fit the general mold by making the identifications $E = B$, $I = C$, and $f = \mathrm{Th}_R$ or Th_M.

The logical conception of representation is attractive since the fundamental idea underlying the notion of logical entailment or derivability is that of identifying the conclusions implicit in given facts. But it does not follow that all

interesting means of identifying implicit conclusions must be forms of logical derivations. In fact, there are strong reasons for thinking that implicit knowledge is, in some cases, both more and less than the deductive consequences of the agent's explicit beliefs, that is, that implicit knowledge can be supralogical or sublogical. These reasons have to do with how the agent handles incomplete knowledge and inconsistent knowledge. 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 or implicit belief since they are not easily stated as aspects of logical theories. For example:

- Some natural categories of implicit conclusions do not follow logically from the explicit knowledge, yet pervade commonsense reasoning. These are recognized as instances of default reasoning or nonmonotonic or circumscriptive inference, but the standard views of implicit knowledge do not know what to make of such nonlogical derivations. If the theory of implicit knowledge is to incorporate such unsound conclusions, the derivation function cannot be purely logical.

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

- Some theories of implicit 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 make some inference on their own, the theory of implicit belief should not ascribe those inferences to the subject. For example, many inferential limitations in artificial intelligence stem from the strategies or procedures which the agent uses to conduct its reasoning. If these procedures avoid drawing some permissible conclusion, perhaps due to limits on available time, memory, or other resources, then the agent's implicit beliefs might well be taken as less than the logical closure of its explicit beliefs, since the logic describes logically

possible inferences, not necessarily economically feasible inferences. In such cases, the implicit beliefs need not be closed under Modus Ponens. (Harman's immediate implications also represent sublogical knowledge, as they are motivated in part to capture such limitations on inferential capabilities.)

- Some of the systems developed in artificial intelligence provide for retracting assumptions by making them defeasible. Defeated assumptions are explicit beliefs omitted from the implicit beliefs upon explicit command.

- Finally, because it insists that the explicit beliefs be consistent, logical theories of implicit knowledge are unable to handle the inconsistent knowledge that arises regularly in artificial intelligence systems. These inconsistencies arise, in the simplest case, because the knowledge of agents is drawn from several experts who disagree about the facts, or who think they agree because the inconsistencies in their views are too subtle to detect. When the agent detects inconsistencies in its explicit beliefs, one response might be to select some consistent subset upon which to reason. In this case, the agent's implicit beliefs might be the consequences of the consistent subset alone, and so omit the remaining, inconsistent explicit beliefs.

In each of these cases, logic alone provides no guidance as to what to do, and there has been considerable debate about how to view these cases of supralogical and sublogical implicit knowledge. One way that has been suggested is to use a nonstandard logic instead of ordinary logic. We may easily get different theories of implicit belief within this framework 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 incomplete ordinary sound rules; Levesque [1984] discusses the use of 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 implicit beliefs. There is no requirement that either explicit or implicit beliefs be complete, but both sets are required to be consistent. In addition, 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 (see especially [Barwise 1985]). The drawback of these deviant logics is that, as with second-order logic, proof procedures do not always exist.

Even though this view of explicit and implicit knowledge is wide enough to incorporate many interesting theories, several problems remain. 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 R or models M. If M must be a subset or superset of the set of ordinary models, this seems unlikely. Even if the framework is completely general in this sense, the resulting logics may not be illuminating, if the logics propose to define a concept rather than reflect an independently known concept. While such logics have the virtue of presenting a precise characterization of a class of implicit conclusions, they do not attempt to explain or understand the nature of the representation relation so captured. What we really seek are conceptually-based 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 satisfying accounts of the nature and suitability of the deviant logics.

The second problem is that some interesting theories do not fit this mold at the outset, as they associate several distinct possible sets of implicit conclusions with each individual set of explicit beliefs. Examples of such logics include McDermott and Doyle's [1980, McDermott 1982] nonmonotonic logics, Reiter's [1980] logic of defaults, and Moore's [1983] autoepistemic logic. 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, for 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 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. To accommodate these ambiguous or nondeterministic sorts of theories, we must change the representation function f to a representation relation or correspondence F, rewriting the suggestive equation above as the condition $I \in F(E)$.

4 Rational representation and reasoning

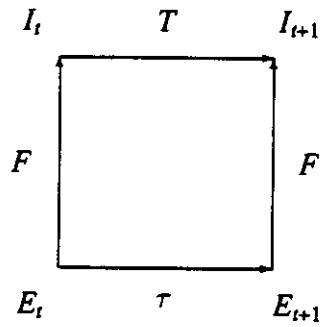
We suggest that a more illuminating theory of implicit knowledge is that it is a case of rational representation, in which the agent treats its explicit knowledge as a specification of its implicit knowledge and rationally chooses how to interpret these specifications to get the implicit knowledge. We employ the standard conception of rationality, in which a choice is rational if it is of maximal expected utility, that is, if the total utility of the consequences of making that choice, discounted by the likelihoods of the consequences of making the choice, equals or exceeds that of any alternative.

If representation involves rational choice, it is easy to see why the relation between explicit and implicit knowledge is a correspondence F rather than a function f. Since rational choice by definition may yield several maximally good possibilities, rather than a single best choice, there may be several sets of implicit conclusions corresponding to a single set of explicit beliefs. As we see below, this phenomenon underlies the nondeterministic theories mentioned above.

To better appreciate this view of implicit knowledge, we must examine the connection between implicit conclusions and inferable conclusions, or more generally, how implicit and explicit knowledge enter into reasoning and action. The first thing to note is that the terms reasoning and inference are widely used in two different senses. One sense is the logical one, the sort referred to in the term "inference rule." In this sense, inferences are proofs, derivations or implications within a formal logical system, structural connections between beliefs or other elements of instantaneous states of the agent. The other sense is the psychological one, in which inference is an activity of the agent. It is clear that logic has little to say about this sense of inference. As Harman [1986] puts it, inference is not implication: reasoning and inference are activities, while proofs in a logic are not activities but atemporal structures of a formal system, distinct from the activity of constructing proofs. Since logical rules of inference tell us nothing about what to do, about what beliefs to adopt or abandon, logic is not, and cannot be, the standard for reasoning. Instead, since reasoning is an activity, the natural standard for reasoning is rationality. Logic, of course, may be employed to formalize psychological theories. For example, logics might be formulated to describe the instantaneous closure and consistency properties of or implications of agent's attitudes, such as the consistency conditions on states

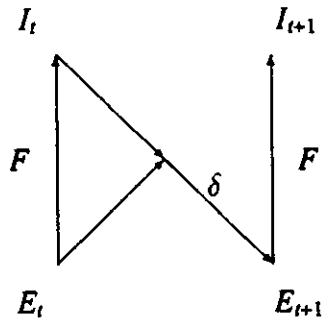
related to rationality, or to axiomatize the possible trajectories of the agent's states. But this use of logic is not particular to psychology, for in the same way logic may be used to formalize meteorology or any other subject matter, and mental operations are not thereby inherently logical operations any more than meteorological events are thereby inherently logical operations.

We can make the distinction between these senses of "inference" more vivid by depicting a step of reasoning as follows.



In this diagram, the vertical arrows indicate the representation relation F, a relation of inference in the logical sense of the connection between the components of the agent's knowledge at an instant. In contrast, the horizontal arrows indicate inference in the psychological sense of actions or changes of state: τ indicating a change in explicit knowledge, and T indicating a change in implicit knowledge. In this paper we will use the terms reasoning and inference only in the psychological sense, and we use representation or derivation for the logical sense. Thus rational inference is just rational conduct of the activity of reasoning, and rational representation is rational choice of sets of derived or implicit knowledge.

The key reason why implicit knowledge is a rational, rather than logical, construct is that reasoning and representation are closely connected in most artificial intelligence systems. Though steps of reasoning may be based on implicit knowledge, the implicit knowledge is changed only indirectly by means of changes in the explicit knowledge. In terms of the diagram, both T and τ are derived relations. A more accurate picture is as follows.



Here we assume for simplicity that the agent's steps of reasoning do not depend on its history or environment, but only on its explicit and implicit knowledge through a function δ. In this case, we have (abusing the notation), $T = F \circ \delta$ and $\tau = \delta \circ F$.

It is natural to view the relations δ, τ, and T as involving rational choice. We might try to keep the theory of representation purely logical by forcing all aspects of choice into δ, leaving F free to be deductive closure, but it is more natural to view F as involving choice as well. The reason for this is simple. In artificial intelligence, δ and F represent computations, not just abstract relations. Each "proper" step of reasoning in the diagram is mechanized by first computing from E_t enough of I_t to allow selection of δ. These computations may be quite separate, involving different procedures; for example, with F involving marker propagation in an inheritance network, and δ involving production rules, conflict resolution, and reason maintenance. But the dividing line is often unclear, since many of the same rules are used in both senses of inference, in computing F and in computing δ. And these rules need not be purely deductive, for nondeductive rules such as defaults are used in both representational and inferential senses (as discussed further below). Thus we can expect computationally important representational relations to involve rational choice as well, since in computational systems the representational relation corresponds to acts of constructing the implicit knowledge by interpreting the explicit knowledge.

The involvement of choice in representation makes the situation in computational reasoning considerably different than the traditional logical conceptions of reasoning in philosophy. Philosophers have long distinguished the notions of theoretical reasoning, which focuses on questions of truth, and practical reasoning, which focuses on questions of value. Since philosophers made no distinction between implicit and explicit knowledge, it has been very tempting to identify

these two dimensions of reasoning with the two dimensions of our diagrams, with implicit knowledge reached by theoretical reasoning and the next state reached by practical reasoning. But this identification is not appropriate, for the distinction between explicit and implicit knowledge is not that between theoretical and practical knowledge. Instead, in computational reasoning representation is also an aspect of practical reasoning, and theoretical reasoning, if it exists at all, is something else entirely, something not necessarily involved in the process of reasoning.

In the following, we examine several ways in which the nonlogicality of implicit knowledge appears to be closely connected with rationality. We first consider the effect of limited inferential or computational resources, and then the effects of incomplete and inconsistent knowledge.

5 Rational expenditure of limited resources

Konolige's [1985] theory of implicit belief attempts to capture limitations on the deductive abilities of the agent by describing implicit beliefs as the closure of the explicit beliefs under a possibly incomplete set of deductive rules. While important for reasons having to do with constitutional reasoning (see [Doyle 1988a]), this view of limited deductive powers is very restricted in the sorts of limitations it can capture. In particular, many limitations may be viewed as arising from using complete sets of rules, but only having a limited amount of time or memory available to draw conclusions. Thus some logics of knowledge attempt to capture limitations on the deductive capabilities of agents by incorporating descriptions of resources into the description of states. That is, instead of describing what implicit beliefs follow from explicit beliefs, these logics describe which implicit beliefs follow from given explicit 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, or that is guaranteed to stay within the expected resource limits.

While theories of knowledge that take limitations and resources into account are a step in the right direction, theories that describe 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 resources are always changing anyway through consumption and possibly through changes in the agent's environment, but 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. Hence there is no natural logic of the limits to reasoning, as these limits are not just a matter of the agent's beliefs. Each logic of limited reasoning (such as Levesque's [1984] logic of explicit and implicit belief, or Davis' [1981] logic of obvious inferences) reflects a fixed set of limits, and no such static logic applies to minds possessed of the ultimate resources of intelligence and industry applied over time.

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. Note that 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. Avoided conclusions are at the heart of defeasible reasoning, underlying many common forms of reasoning and representation, yet the very idea of avoided conclusions contradicts the implicit assumption of the logical view of representation that knowing more is always better. Deliberate ignorance is foreign to scientists, who are trained to want to know everything. But it is common in the everyday lives of people who often would rather not know something they could easily find out. This is not merely a human foible. Jonathan Cave pointed out to me that game theory has studied numerous examples of this phenomenon, and he kindly provided me with the following illustration. (See also [Cave 1983]. We will not explain the terminology of game theory here. See, for example, [Luce and Raiffa 1957].) Suppose that two players are to play a 2×2 game simultaneously and once only. The real game may be game A or game B below.

8, 8	0, 10
10, 0	2, 2

Game A

8, 8	4, 4
4, 4	0, 0

Game B

8, 8	2, 7
7, 2	1, 1

Average

Game A is prisoners' dilemma, where cooperation is best but betrayal is individually rational. It has a unique dominant strategy equilibrium with payoff (2, 2). Game B is a game of coincident interests in which cooperation is individually rational, and has a unique dominant strategy equilibrium with payoff (8, 8). Suppose that the prior probability of each game is 1/2. If the players are informed of the true game, they will play its unique equilibrium, for an overall expected payoff of $(5, 5) = .5(2, 2) + .5(8, 8)$. If they are not informed of the true game, their expectations are described by the "average game" matrix above. This game has a unique dominant strategy equilibrium with payoff (8,8). Thus even if the players can find out which situation actually obtains, they are better off not knowing, since if the actual situation is game A, they will betray each other and be much worse off. In this example, it is the unnecessary ignorance that makes beneficial cooperation rational. In this case, ignorance is most certainly bliss.

The rational view of representation suggests that the underlying source of these difficulties is that resource-limited reasoning, as it is called, is an incomplete idea. The knowledge available to or exhibited in action by the agent depends on its preferences as well as on its beliefs and resources. 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. Thus 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. And since its preferences may depend on its plans, desires, and other attitudes, the agent's knowledge is determined by all of the agent's attitudes, not just its beliefs and strictly computational resources. Resource-limited reasoning is really a code-word for the economics of reasoning, for the rational allocation of resources. But extant suggestions about the effect on representation of resource-limited reasoning focus on cases in which the agent is bound to draw every conclusion it can within the limits of its resources. The more natural and general view is that in rational representation there may be several sets of implicit conclusions corresponding to a single set

of explicit beliefs, each representing a different allocation of resources, possibly with little overlap between the distinct choices.

6 Rational assumptions and incomplete knowledge

As we just saw, rational choices influence implicit knowledge even when the explicit knowledge is complete. We now consider specifically the case in which the explicit knowledge is incomplete in ways relevant to the agent's actions. Here rational choice enters through assumptions included in the implicit knowledge at the behest of rules of assumption included in the explicit knowledge.

The logical view of representation offers no means for explicit knowledge to indicate implicit assumptions, even though thinking often begins with making guesses grounded in one's experience. But even though guessing, or making assumptions, is often held in disrepute as illogical, it is often quite the rational thing to do. 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. But the first limitation faced in limited reasoning is that one cannot either know or consider everything, and so must 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. Most of the cases of default reasoning appearing in artificial intelligence represent judgments that, in each particular case, it is easier to make an informed guess and often be right than to remain agnostic and work to gather the information; that errors will be easily correctable and ultimately inconsequential; and that the true information needed to correct or verify these guesses may well become available later anyway in the ordinary course of things. 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.

Rationality may be applied as a standard motivating the adoption of individual defaults in a very natural way, by saying that an assumption or rule of assumption should be adopted if the expected utility of holding it exceeds the expected utility of not holding it. Applied to individual assumptions, this is a familiar idea, famous under the names of Pascal's wager in the case of religious belief, and James' "will to believe" for the general case of religious and mundane beliefs. For example, Pascal [1662] framed his problem of belief in God as the following: 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 a decision matrix

Pascal's decision	God exists	doesn't
Believe	$+\infty$	$-f$
Doubt	$-\infty$	$+f$

where f represents the finite pleasures enjoyed or forgone during Pascal's life. Of course, these same quantities modify the first column as well, but finite modifications to infinities 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, $-\infty$. Note that this evaluation of possible beliefs means taking both utility and probability into account. It is not 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. But it is also not 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 1987]). This mistake has no notorious name, and perhaps not coincidentally, has enjoyed some popularity in artificial intelligence. When both probability and utility are taken into account, it is in some cases rational to make assumptions expected to be false, just as we saw that ignorance is sometimes rational. For example, when asking directions

on an army base, if one cannot read insignias of rank it is advisable to assume all soldiers have high rank (such as colonel). Since there many more soldiers of low rank than of high rank, one expects this assumption to be false. But one consequence of error is to have the soldier give his true rank, and the consequence of the expected error is to flatter the soldier, making him more eager to help. 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.

In artificial intelligence, rules for making default assumptions have been at issue more than individual assumptions, though the two cases may be assimilated since individual default rules may be evaluated as individual assumptions. This view of adoption of default rules has been urged by [Doyle 1983] and [Shoham 1987]. Once adopted into the explicit knowledge, the rules may be applied to produce assumptions either when needed during construction of implicit knowledge, as in most inheritance systems, or in advance as additional parts of the explicit knowledge, as in reason maintenance.

7 Rational interpretations of inconsistent knowledge

Another limitation of the logical view of implicit knowledge is its insistence on consistency in the explicit knowledge. Although there are some logics intended to permit reasoning from inconsistent knowledge, none of these are very compelling, and few offer the sort of independent justification for their structure we sought in the case of unsound inferences. Here also the idea of rational inference offers an approach to representation in the presence of inconsistent knowledge. In this approach, the agent rationally chooses a consistent subset of its explicit knowledge, and then uses this subset to choose a consistent body of implicit knowledge (though these need not be separate decisions since the subset may be chosen so as to yield a desired conclusion). In such cases we may think of the selected implicit knowledge as representing the inconsistent explicit knowledge for the purpose of the action at hand, with the agent possibly selecting different representations of the explicit knowledge for subsequent actions. For example, the main theories in artificial intelligence of reasoning with inconsistent knowledge are those exemplified by reason maintenance, nonmonotonic logic, and the logic of defaults. Appropriately reformulated, these are all approaches

towards reasoning with inconsistent preferences about beliefs. Specifically, the nonmonotonic justifications of reason maintenance, the nonmonotonic implications of nonmonotonic logic, and the default rules of the logic of defaults are all better viewed as expressing preferences of the agent about what conclusions it should draw. (See [Doyle 1983, 1985, 1988a,c]. [Etherington 1987] makes a similar suggestion.) Each of these theories sets out possible sets of conclusions which correspond to choosing conclusions on the basis of certain (in particular, Pareto-optimal) consistent subsets of the inconsistent preferences. Of course, each of these theories takes into account only very simple sorts of preferences, and constructs mental states that are rational with respect to the special classes of preferences but which may be irrational with respect to the classes of more complicated attitudes ignored in the construction.

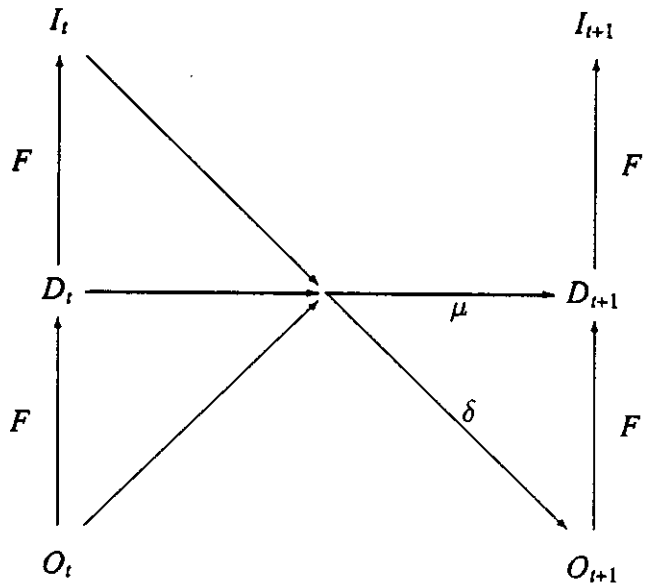
In the formalization of [Doyle 1988a], the problem of acting with inconsistent knowledge is formally identical to the problem of group decisions or social or political action when the members of the group conflict, justifying our use of the term "representation" for these choices. (See also [Minsky 1986].) This means that the whole range of techniques for making decisions in the presence of conflict studied in politics and economics may be adapted for use in the case of inconsistent individual action. Correspondingly, architectures developed in artificial intelligence might be considered as possible structures for human governments. But in each case, the motivations and merits of an organization must be re-evaluated in its new setting. For instance, the traditional approach toward inconsistency in artificial intelligence has been to abandon some of the inconsistent explicit knowledge by replacing the inconsistent set with the selected consistent set. In politics, this is just the ancient technique of killing (or at least exiling) one's opponents, a technique no longer countenanced in democratic states. In this setting, the "clash of intuitions" about inheritance reasoning observed by Touretzky, Horty, and Thomason [1987] is a special instance of the larger difficulty of satisfying, in one form of government, all reasonable desiderata for governments. See [Doyle 1988c] for more on this.

8 Mechanizing rational representation

Even if rationality provides a more appropriate ideal standard for representation and reasoning than logicality, as a practical matter it poses difficult problems,

problems at least as difficult as those faced in mechanizing logical reasoning, and problems of the same character as those faced by human thinkers that rationality in reasoning seeks to address. For example, often our knowledge of the costs and benefits of assumptions is incomplete and itself consists mainly of guesses. We can of course look for selections rational with respect to these guesses, but trial and error is the rule with these cases, so we call them heuristics rather than calculated assumptions. In most current artificial intelligence systems, these judgments or calculations are made by the system's designer—the human informants decide what the good guesses are, and these are encoded into the rules that the machine obeys. Alternatively, these judgments might be made by the agent itself as well through reflection and reasoning. But going further, for machines as well as people it is too hard, in general, to explicitly compare all possible states of knowledge, for the states may be infinite in both size and number. One may not then be able to choose states rationally simply by retrieving one's expectations and preferences and choosing the state of maximal expected utility. This does not mean that the idea of rationality is useless in mechanizing reasoning, for its role as the ideal theory of reasoning is important even if it cannot be fully attained. Instead, one must find ways of approximating rational representation and reasoning by using simple elements of rational information in ways compatible with their intended meanings. For instance, since the agent will only be able to expend limited effort on reasoning decisions, it might allocate this effort by another rational choice involving restricted sorts of preferences, such as those expressed in default rules. In this setting, the dialectical and successively reflective patterns of preferences about preferences and meta-level reasoning described by [Doyle 1980] and [Smith 1985] may be viewed as elements of a theory of rationally bounded rationality (see [Doyle 1988a]).

In fact, several important notions in artificial intelligence systems can be directly interpreted as computational approximations to the distinction between explicit and implicit knowledge. These include Levesque's [1984] notions of explicit and implicit knowledge and the notions of derivative knowledge and conservative revisions as they appear in reason maintenance. In Levesque's theory, a subset of the implicit knowledge is accepted as effectively "explicit" knowledge because it is easily computable. In reason maintenance [Doyle 1979], we have the situation depicted in the following diagram.



Here the explicit knowledge $E_t = O_t \cup D_t$ is divided into two parts: an "original" part O_t and a "derivative" part D_t, where as before, explicit changes are made only to the original knowledge via δ, and in addition the system updates the derivative knowledge as necessary via μ. This original/derivative distinction reflects the explicit/implicit distinction within the explicit knowledge itself, since if we relabel the original knowledge as "explicit," then the derivative knowledge is part of the "implicit" knowledge corresponding to the original knowledge. That is, we may have $D_t \subseteq I_t \in F(O_t)$ in addition to $I_t \in F(O_t \cup D_t)$. Moreover, the revision process serves as a way of making the construction of "implicit" (derivative) knowledge more efficient (in intent anyway, if not in practice) by recording and revising earlier portions of the implicit knowledge for use with modified bodies of "explicit" (original) knowledge. These records are then used to effect conservative revisions of the derivative knowledge, in which the new state D_{t+1} is chosen to be as "close" to D_t as possible among the alternatives in $F(O_{t+1})$. (See [Doyle 1988a,d] for more on conservatism and rationality.)

9 Conclusion

To summarize, implicit knowledge is an essentially decision-theoretic notion, not a logical notion, and the limits to knowledge are primarily economic, not

logical. The agent's implicit knowledge depends upon its preferences as well as its beliefs, with these preferences changing over time. This means that no static logic of belief (or even of belief and resources) can capture notions of implicit belief conforming to commonsense ascriptions of belief.

The nonlogical nature of implicit knowledge is less surprising when considered from the larger perspective in which representation is just one aspect of the organization of reasoning, for there it is easier to see immediately the limitations of logic as a set of principles of reasoning. What is lacking in logic as even an ideal theory of thinking is that reasoning has a purpose, and that purpose 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, discarding old ways and inventing and adopting new ones. Guesses, rational or not, are logically unsound, and instead of preserving truth, reasoning revisions destroy and abandon 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 right or 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 the others, whether true or false. Even though one might hope to organize representation and reasoning to avoid non-logical assumptions and revisions, to instead involve only cumulative, logical deduction of the consequences of initial knowledge and descriptions of passing experiences, it hardly seems possible to live that way. Guesses are necessary, for humans at least, because of the frailty and smallness of human mental capacities. 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. Revisions and reinterpretations of knowledge in turn are necessary because even if guesses are never wrong, progress in reasoning, like maturity and progress in life, requires escape from the shackles of the past. Agents whose knowledge is cumulative, being unwilling to give up the past, are condemned to repeat it endlessly. Put most starkly, reasoning aims at increasing our understanding; rules of logic the exact opposite.

Acknowledgments

An abbreviated predecessor 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, and this paper itself abbreviates some of the material contained in a much longer work, [Doyle 1988a].

I am especially indebted to Hector Levesque, Ronald Loui, and Robert Moore for their commentaries ([Levesque 1988], [Loui 1988], and [Moore 1988]), to Jonathan Cave for the example of rational ignorance (I have used portions of his explanation verbatim), and to Allen Newell, Jonathan Pollock, Joseph Schatz, Richmond Thomason, and Michael Wellman for valuable comments and ideas.

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, *Seventh IJCAI*, 530-531.

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, *Proc. Second Conf. on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 345-354.

Doyle, J., 1988c. On universal theories of defaults, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-111.

Doyle, J., 1988d. Similarity, conservatism, and rationality, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-123.

Etherington, D. W., 1987. A semantics for default logic, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 495-498.

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

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

Levesque, H. J., 1984. A logic of implicit and explicit belief, *AAAI-84*, 198-202.

Levesque, H. J., 1988. Comments on "Knowledge, representation, and rational self-government," *Proc. Second Conf. 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, *Proc. Second Conf. on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 355-359.

Luce, R. D., and Raiffa, H., 1957. *Games and Decisions*, New York: Wiley.

McDermott, D., 1982. Nonmonotonic logic II: nonmonotonic modal theories, *J. A. C. M.* **29**, 33-57.

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

Minsky, M., 1986. *The Society of Mind*, New York: Simon and Schuster.

Moore, R. C., 1983. Semantical considerations on nonmonotonic logic, *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?, *Proc. Second Conf. on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann, 363.

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

Pearl, J., 1987. Probabilistic semantics for inheritance hierarchies with exceptions, Los Angeles: UCLA Cognitive Systems Laboratory, TR-93.

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

Shoham, Y., 1987. Nonmonotonic logics: meaning and utility, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 388-393.

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

Touretzky, D., Horty, J., and Thomason, R., 1987. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems, *Ninth International Joint Conference on Artificial Intelligence.*