KNOWLEDGE, REPRESENTATION, AND RATIONAL SELF-GOVERNMENT

(Position Paper)

Jon Doyle
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

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 nonlogical inferences is that they are cases of rational inference, in which the agent rationally (in the sense of decision theory) chooses the conclusions it wishes to adopt. Thus in rational inference, the implicit beliefs 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 beliefs themselves. The explicit representations possessed by the agent are not viewed as knowledge themselves, 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 the cases omitted by the narrow logical view, subsuming and explaining many AI approaches toward reasoning with incomplete and inconsistent knowledge.

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 described by a logical theory of thinking that sets out the laws of thought and principles of reasoning. We suggest an alternative conception of knowledge, based on the notion of rational inference, that overcomes these limitations in natural ways. In particular, deviations from logicality are conventionally viewed as "performance" failures that do not reflect upon the suitability of the logical "competence" theory. In the theory of rational inference, the common sorts of deviations from logicality are seen to be part of the competence theory, not mere failures in performance.

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. 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

$$K_I = f(K_E),$$

where K_I stands for the agent's implicit knowledge, K_E for the agent's explicit knowledge, and f for the function describing how explicit knowledge determines implicit knowledge.

LOGICAL AND NONLOGICAL REPRESENTATION

In fact, the standard view of explicit and implicit knowledge focuses on explicit and implicit belief, following a long tradition in philosophy that views knowledge as something like true belief (see, for example, [Moore 1985]). According to this view, we would write

$$C = \mathrm{Th}_R(B)$$

indicating that the base beliefs B represent the conclusions C via closure under a set of sound deduction rules R (see, for example, [Konolige 1985]). This theory clearly fits the general mold. In this way, each epistemic or doxastic logic describes a theory of knowledge or belief, and the theories of knowledge in the literature differ mainly in the logic used, as captured in the closure function Th_R. Such theories have very nice properties: the agent's beliefs may be incomplete, but most logics considered are sound, so that the implicit beliefs are consistent when the explicit ones are. For example, Moore [1985] employs standard epistemic modal logics; Konolige [1985] permits incomplete ordinary sound rules; and Levesque [1984] argues that the logic should be relevance logic.

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, and the derivation rules are required to be sound, that is, truth preserving according to ordinary models. This does not mean that the logic must be an ordinary one, for if the set of models under consideration is set out independently, the logic has embedded concepts, and may have important nonstandard characteristics (see [Barwise 1985]).

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. These reasons have to do with handling incomplete knowledge and inconsistent knowledge. Most of these examples and reasons are fairly well known, but have not been fully incorporated into theories of knowledge or implicit belief since most are not easily cast as logical theories. In each of these cases, logic alone provides no standards for what to do. There has therefore been considerable debate about how to view these nonlogical inferences.

SUPRALOGICAL KNOWLEDGE

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 rule cannot be purely logical.

Harman's [1986] notion of immediate implication provides another example of supralogical 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.

SUBLOGICAL KNOWLEDGE

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 are motivated in part to capture such limitations on inferential capabilities. As another example, 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.

RATIONAL REPRESENTATION

Rather than view them as failures of logic, we view the various sorts of nonlogical inference as cases of rational inference. For our purposes here, 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. 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.

RATIONAL ASSUMPTIONS

Thinking often begins with making guesses grounded in one's experience. Guessing, or making assumptions, is often held in disrepute as illogical. In fact, though 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. To minimize or avoid these 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 so these savings must be weighed against the expected costs of making errors (see, for example, [Langlotz et al. 1986]). 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, assumptions are

avoided, either because there is no information available to inform the guess, or because even temporary errors of judgment are considered dangerous.

RATIONAL CONTROL OF REASONING

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. While theories of knowledge that take resources into account are a step in the right direction, they have several limitations. The first difficulty is that choosing the wrong notion of resources yields an uninteresting theory. The second difficulty is that the quantities of resources available to the agent need not be well defined, since 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. The third difficulty is that the agent may have the resources to draw a conclusion, but no interest in (or even a definite antipathy for) drawing it. Note that these limitations are not simply a matter of competence and performance, for we would think an agent incompetent if it could not avoid actions it intends to avoid and has the power to avoid.

The rational view of reasoning suggests that the underlying source of the second and third difficulties is that resource-limited reasoning, as it is called, is an incomplete idea. The knowledge available to or exhibited by the agent in action 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. Resource-limited reasoning is really a code-word for the economics of reasoning, for the rational allocation of resources, but extant suggestions about resource-limited reasoning which focus on cases in which the agent is bound to draw every conclusion it can within the limits of its resources. In contrast, in rational inference there may be several sets of implicit conclusions corresponding to a single set of explicit beliefs, each representing a different rational choice, possibly with little overlap between the distinct choices. 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 merely its beliefs and merely computational resources. 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 or approximations to rational inference, in which the effort applied in rationally choosing inferences is allocated by another rational choice involving restricted sorts of preferences. This leads into a theory of *rationally bounded rationality* which we cannot pursue here (see [Doyle 1987]).

RATIONAL SELF-GOVERNMENT

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 the idea of rational inference offers, with no added notions, an approach to reasoning from inconsistent knowledge. In this approach, the agent rationally chooses a consistent subset of its explicit beliefs, and then uses this subset to choose a consistent set of implicit beliefs (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, possibly selecting different representations of the explicit knowledge for subsequent actions. For example, in artificial intelligence, the main theories of reasoning with inconsistent knowledge are those exemplified by reason maintenance, nonmonotonic logic, and the logic of defaults. Appropriately reformulated (see [Doyle 1983, 1985, 1987]), 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. 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.

In the formalization of [Doyle 1987], 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. 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 (see, for example, [Minsky 1986] and [Wellman 1986]). 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 et al. 1987] is a special instance of the larger difficulty of satisfying, in one form of government, all reasonable desiderata for governments.

CONCLUSION

To summarize, implicit knowledge is an essentially economic 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. 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. 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. Put most starkly, reasoning aims at increasing our understanding; rules of logic the exact opposite.

Acknowledgments

This paper abbreviates some of the material contained in a much longer work, [Doyle 1987]. I thank Joseph Schatz, Richmond Thomason, Michael Wellman, and Allen Newell for valuable comments and ideas. This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 19, 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.

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.

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., 1987. Artificial intelligence and rational self-government, Pittsburgh: Carnegie Mellon University, Computer Science Department.

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.

Langlotz, C. P., Shortliffe, E. H., and Fagan, L. M., 1986. Using decision theory to justify heuristics, *Proc. Fifth National Conference on Artificial Intelligence*, 215-219.

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

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

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.

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.*

Wellman, M. P., 1986. Consistent preferences in a mind society, MIT 6.868 term project report, unpublished.