The role(s) of belief in AI

Conference Paper · December 2000

DOI: 10.1007/978-1-4615-1567-8_16

1 author:

Don Perlis
University of Maryland, College Park
36 PUBLICATIONS **82** CITATIONS

SEE PROFILE

Chapter 16

THE ROLE(S) OF BELIEF IN AI

Don Perlis*
University of Maryland
Institute for Advanced Computer Studies
and Department of Computer Science
College Park MD 20742
www.cs.umd.edu/users/perlis
perlis@cs.umd.edu

Abstract Beliefs play complex and sometimes confusing roles in AI. This paper surveys (i) a variety of notions of belief, (ii) formal efforts to characterize beliefs, and (iii) how beliefs are related to action, to language, and to commonsense. In addition, we will consider certain logical tensions between beliefs and consistency.

Keywords: Belief, consistency, modality, content

1 INTRODUCTION

Already by 1980 many researchers in AI were hard at work on the problem of belief, and philosophers had been at work on it even before that. But then (Nilsson, 1983) introduced the idea of robots with a lifetime of their own. This notion, which we now refer to more generally as autonomous intelligent agents, makes the role of belief central, for such an agent will have to behave in ways only explainable in terms involving belief, as will be argued below.

To clarify one point at the outset: we use the term "belief" here *not* in reference to a mere suspicion or impression that the item believed is possibly true, as in "I believe so"; rather it is in the much stronger sense: what the believing agent takes to be true, as if it were possible to look inside the agent's "head" and see what the agent's view of the world is. And of course, with artificial agents, it is so possible; but this leaves open the problem of which items found there are to count as beliefs (or as views of the world).

*The author would like to express thanks for support for this research, from grants funded by AFOSR and ONR.

This paper is organized as follows: Section 2 describes and contrasts a variety of notions of belief in AI; Section 3 discusses why beliefs are essential to the AI enterprise; Section 4 examines the extent to which beliefs are formal (or inference-bound), and Section 5 considers certain tensions between beliefs and consistency.

It may be that someday we will succeed in building an artificial autonomous agent that, by all routine standards, is intelligent. And yet if we do not understand how that agent works, if we cannot explain its behavior in a way that sheds light on the nature of intelligence, then we shall have succeeded as engineers but not as scientists.[1]

One key component of the scientific enterprise of artificial intelligence is the understanding of the nature of beliefs. This paper discusses that component.

2 WHAT ARE BELIEFS?

What constitutes a belief is controversial. A number of distinct characterizations have been offered, serving best perhaps to underscore the notion that this common term of everyday usage masks a number of highly distinct phenomena. Early on there appears to have been a conflation of belief, knowledge, and data, as in the expressions "knowledge base" and "belief base" used in reference to a "database" of "facts". In addition, much early work identified an agent's beliefs with inferential consequences of its database.[2]

From a programmer's or logician's point of view, a belief can be thought of simply as a piece of data that has a truth value (it is true or false), and that can in some form be available to an agent to conduct its reasoning, much like an axiom or theorem. But in more cognitive terms, an agent's belief base can be thought of as its view of what the world is like. These two encompass quite a range of differences, and each of them admits of further distinctions as well. For instance, axioms are sentences in a formal language; yet beliefs are sometimes taken by the logic community to be abstract propositions rather than sentences. This will be discussed below in terms of modal versus sentential representations.

Partly in an effort to lend some clarity to this confusion, (Levesque, 1984) has emphasized the difference between implicit and explicit belief: the former includes all consequences of the agent's database, and the latter only those that the agent has in fact obtained. The latter of course can (and usually does) change over time.[3]

---

[1] It is said that Henry Bessemer, who invented the Bessemer process for making steel, had no principled understanding of how his process worked, but that he tried many approaches until one finally did the job. Thus the above remark should in no way be read as disparaging of engineering; it is simply a matter of differing aims.

[2] (Gettier, 1963) has shown that knowledge is a very tricky notion, and many have simply given up on knowledge as a useful notion at all, in favor of belief. However, belief is not without its own pitfalls as we shall see.

[3] There is also (see Section 5) a notion of *evolving* explicit belief (formalized in terms of so-called active logics) which characterizes a reasoning process by which beliefs can be formed, refined, and rejected.

Levesque's *implicit* version is often stated in terms of the property of *logical omniscience*: such beliefs are closed under logical consequence. Thus if α and $\alpha \rightarrow \beta$ are agent beliefs, so is β, even if the agent in question has not in fact actually drawn that conclusion. One says the agent implicitly believes β. See (Fagin et al., 1995) for a detailed recent study of this approach.

(Konolige, 1986) has described in detail how an actual agent might construct beliefs from one beliefs in a *deductive* process akin to Levesque's explicit beliefs, forming thereby the set of all such inferred beliefs. For Konolige this process may be highly constrained due to resource limitations. On this approach it is possible to believe α and $\alpha \rightarrow \beta$ and yet fail to deduce β, thereby also failing to believe β. On the other hand, Konolige's treatment does require the belief set to be closed under all inferences available to the agent, a limited form of omniscience: the agent believes all it ever comes to believe. One can also consider a (usually) larger class, Deduct*, of those beliefs that would arise from given inferential processes without any resource bounds.[4]

It may be the case that explicit beliefs coincide with deductive beliefs; this is so if we consider the believing agent only at the *end* of its reasoning efforts, when it has come to believe all it will believe. And it may be that implicit beliefs coincide with the above class Deduct*; this is so if the inference rules are complete. However, all the above notions of belief presuppose some starting set of beliefs, and from those a wider set of consequent beliefs may be characterized.

Another view is that encodings of statements of purported fact, whether in a database or not, cannot be properly viewed as beliefs unless they are connected suitably to the agent's behavior. Agents can spin conclusions out of propositions as mere exercises, just as we can conclude from A and from A implies B, that B, without our believing A, B, or A implies B. Somehow a belief seems tied to our overall behavior and not just to our ability to spin conclusions: we need to be able to believe the premises, as well as the conclusion. This however is no definition, for it is circular. Even putting A in a database is little help, for one can easily have a list of disbelieved propositions or sentences (e.g., the Moon is made of green cheese, babies are brought by storks, etc) all

[4] There is however a curious phenomenon about resource-bounded belief: the very fact of boundedness can lead to beliefs that would *not* arise in the absence of those bounds. This is most notably so when there are reflective inference rules, i.e., ones that produce inferred beliefs reflecting past behavior including absences of certain potential beliefs. For instance, if $NotFlies(tweety)$ is a logical consequence of given axioms but is not derived due to time or space limitations, $\neg Know\ NotFlies(tweety)$ may be inferred. An unbounded (ideal) reasoner would not reach this conclusion but rather $NotFlies(tweety)$ instead. A more dramatic case is that of inconsistent beliefs, which in the case of a resource-bounded reasoner may be limited in the damage they cause, but an unbounded reasoner will produce the usual disaster of *ex contradictione quodlibet*: everything follows from a contradiction. Finally, a temporal resource boundedness can, in the presence of suitable reflective rules, lead to conclusions such as $Now(i)$ and even $Contra(i, P, \neg P)$ to note the time is now i and that a contradiction has been discovered at that time. Unbounded reasoning as it is usually understood in AI does not take account of the passage of time during reasoning and thus has no way to track this evolving character of beliefs over time. See (Elgot-Drapkin and Perlis, 1990; Elgot-Drapkin et al., 1993).

kept in a database. Nor is it enough that such a database be operated on by inferences, as we just saw.

This suggests a possible "dispositional" definition as follows (Perlis, 1986): an expression E occurring in an agent's database(s) is a belief if the agent is disposed to act on E.[5] On this view of beliefs, which I call *use-beliefs*, actions are highly dependent on beliefs. I suspect this is the notion of belief that is most closely relevant to commonsense reasoning; reasons will be given below in Section 3.

This is distinct from but related to, another dispositional account: agent G believes E if G is disposed to *assent* to E if asked "Do you believe E?", or "Is E true?" We may then say G has assent-belief E.

Thus we have seen at least four distinct notions of belief:

- Entailment-belief: this is Levesque's notion of implicit belief: those wffs (or propositions) that are entailed by given (starting) beliefs.
- Deductive-belief: Konolige's notion of belief as those wffs that are proven by the agent in question.
- Use-belief: the notion of a belief as an item (wff or proposition) the agent is disposed to use in making decisions as to what actions to take (in conjunction of course with whatever desires and preferences the agent may have).
- Assent-belief: the notion of a belief as an item the agent is disposed to assent to when asked.

The first two do not say what it is to be a belief, but rather what it is for a belief to arise from other beliefs. The latter two attempt to define belief in more primitive terms. It is unclear how we are to assess whether an agent believes any given item, on most of these notions (assent belief may be an exception); and of course these distinct notions do not agree among themselves in general.[6]

The entailment-belief notion above leads to elegant mathematical properties, but seems of little practical relevance for physically realizable agents. Unfortunately it tends to be forced by most modal treatments which take the items of belief to be abstract propositions of some kind, rather than patterns such as sentences that can be stored in a physical medium like a computer database or a brain. This latter dichotomy can be put differently: are beliefs syntactic objects (such as sentences) in a representational format, or are they more abstract "meanings"? In modal logic beliefs are objects of modalities, i.e., of operators that apply to abstract propositions. For instance one may write *Bel P* where *P* is not an argument to a predicate *Bel*; here instead *Bel*

[5] To be disposed to do X is to have the tendency to do X, in the absence of exceptional circumstances; i.e., X is the default action.

[6] A referee has pointed out the following helpful example to distinguish the latter two: a native Chinese speaker who believes that it is raining may not assent to "Is it raining?" due to not knowing English; but he will have be able to use the belief in order to decide to take an umbrella. This is related to a puzzle discussed later under the heading of abstract content.

is a kind of modifier to P similar in spirit to a connective or negation symbol, telling us in what "mode" or "attitude" to take P, i.e., *Bel* tells us to take it not as true or false or possible or likely or hoped-for, but as believed (by whatever agent is understood). P itself is whatever the sentence "P" represents, which is usually taken to be an abstraction, such as the set of all wffs with the same truth value as the sentence "P".

The above modal approach, due initially to (Hintikka, 1962), is in stark contradistinction to the "sentential" view, namely that beliefs simply are sentences (in some special relation to the agent that believes them, of course, whether assentual or implicit or deductive or use). As AI moves closer to Nilsson's notion of "robots with a lifetime of their own," we will need to be much clearer about such things. As pointed out above, the mere fact of having inferred something in no way leads to willingness to use that conclusion to make and carry out plans. At the very least this depends on how we regard the premises of the inference. So when one robot reasons about another robot's beliefs, something will be occurring that we do not as yet understand very well, and that therefore we are ill-prepared to design. Is the second robot viewed by the first robot as simply having inferred various conclusions? Or as taking a more active stand toward those conclusions? Moreover, two agents (people, robots) are unlikely to use identical representations for their beliefs, so that if B has the belief b, and C comes to believe that B so believes, C's belief may well have to represent B's belief in terms C can understand, which may not be the way B represents it. This is in part the problem of belief-ascription (Kripke, 1979), about which more below.

Yet another – very radical – view is that beliefs are an illusion of naive "folk psychology", and no such phenomenon as believing actually occurs; nor indeed will any more precise counterpart arise to take the place of a currently ambiguous and poorly-formed notion of belief. This view is called "eliminative materialism"; (Churchland, 1984) has presented a detailed defense of this view.

3 WHY ARE BELIEFS IMPORTANT?

Let us call the view that intelligence involves in an essential way the forming of beliefs – i.e., manipulable informational units that obey some sort of semantic strictures – the belief-theory. Much of the AI literature then at least tacitly assumes the belief-theory. An alternative view is that intelligent behavior can and does go on even in the absence of beliefs about the very world that the intelligent agent deals with. Thus neural networks are sometimes offered as examples of intelligent systems that can be made sense of without the belief-theory; this is the above view espoused by Churchland, for instance.

Yet I will argue now that the only account that has been given of intelligent behavior so far, and likely the only one possible, is based on the belief-theory, and in particular on use-belief as the most fundamental notion of belief.

Imagine an agent, Susie Q, who appears to be intelligent. She exhibits a range of behaviors highly sensitive to details of her surroundings, including details of interactions with other agents. In particular, she regularly plays tennis with a partner, Freddy, by advance arrangement with him.

Now, to explain why it is that Susie Q goes to the gymnasium on Tuesday evenings, one might suppose that she wants to play tennis, and that she believes she can do so at the gym, that today is Tuesday, that the gym is open Tuesday evenings, and that Freddy will be there. Moreover, one supposes that were she to be told (a) the gymnasium is closed for repairs, she will not go there; and that if she were told (b) of a new policy that gymnasium members must now bring their own towels, she will either bring a towel or will not go at all; and so on.[7]

No account of these and similar behavioral patterns has ever been given, that does not rely profoundly on the formation and use of beliefs on the part of the agents in question.[8]

One cannot escape this by appealing to a distributed representation such as a neural network: a neural net model may well account for Susie's recognizing (or even learning to recognize) the gym when it is in front of her, or for that matter recognizing when she is in need of exercise, and even that it is Tuesday. But it will not account for her ability to rapidly shift her behavior on being told either of the new statements (a) and (b) above. These latter require her to process her recognitions along with (a) or (b) in ways that are deeply sensitive to details of just the sort that beliefs are thought to have: propositional content – they can be true, false, composed of simpler beliefs, refashioned, and so on; i.e., they have an essentially symbolic or language-like character. While it is perfectly true that, in principle, a neural network or other distributed representation can be designed so as to provide this sensitivity, such a design amounts precisely to turning that representation into a belief-system! That is, it will have to process informational units according to semantic strictures such as modus ponens.

Thus it is not only the case that until someone comes up with a different and equally explanatory account of such behaviors, we are stuck with the belief-theory. The case is stronger: the behaviors we wish to account for, and to design into agents, are themselves ones of semantically-bound processing of informational units. The belief-theory appears to be true on the basis of the very nature of the problem. And therefore, if this is so, beliefs are fundamental to the AI enterprise; the entire field, one might say, amounts to the learning, inference, recall, use, communication, and manipulation of semantically bound informational units, whether in planning, scene interpretation, natural language processing, problem solving, navigation. In saying these are semantically bound, we are saying that they are sensitive in very particular ways. We turn to this below.

[7] Note that an essential ingredient in Susie's actions is that she *wants* something; it is for that reason that actual behaviors arise out of the reasoning. That is, it is not beliefs in inference rules alone, but these in conjunction with a pattern of desires and preferences, that leads to the observed behavior.

[8] There are actually two types of behavioral pattern in the above description: those behaviors ascribed to Susie, and the suppositional behaviors ascribed to us who think about Susie. Both involve belief.

4 WHY ARE BELIEFS (PARTLY) LOGICAL IN NATURE?

As mentioned above, beliefs have language-like character: they belong to a more general sort of entity that can be affirmed, denied, questioned, composed into more complex forms, and so on. Moreover, they can be used, as in the Susie Q example, to create new beliefs according to precise rules that are deeply sensitive to local formatting details, such as negations. Thus when told that the gym is not open today, she must process this as a kind of denial of her former belief that it is open. The "not" is more than a mere additional set of pixel data enriching an image; it has specific content that bears on the meaning (and hence the use) of the whole informational unit, saying in effect that the rest of that unit is not to be used.

But these are precisely what characterize a logic: precise rules of composition and formation (this is the "syntax" half) and semantics, where beliefs also do their part: as language-like entities they can have meanings, so that when Susie comes to believe that her regular gym is not open, her actions (staying home, or going to another gym) fit well into the external world. This is due to a good match between her beliefs and the world, and this match is given by the semantic content (meaning) of her beliefs. Of course, beliefs are often wrong, but this too depends on beliefs having semantic content.

Now, none of this is to say that the kinds of beliefs people (or autonomous agents in general) may come up with need follow some standard textbook inference rules. Not at all. Indeed a large part of theoretical AI is devoted to the design of new formalisms that come closer to the needs of real agents. And that is where many of the most interesting and important developments have been and, hopefully, will be: what sorts of belief-manipulations (inferences, putting beliefs together to come up with new beliefs) are appropriate for an autonomous agent? Non-monotonicity is one such development (see (Ginsberg, 1987)); probabilistic reasoning is another, recently shown to be related to nonmonotonic reasoning (see (Bacchus et al., 1993)).

5 WHY ARE BELIEFS PROBLEMATIC?

We have already seen that beliefs are complicated and controversial. But there is worse in store.

5.1 BELIEFS CAN BE INCONSISTENT

(Montague, 1963) showed that, in the case of knowledge, apparently mild assumptions about an agent's knowledge *must* be inconsistent. This was quite a surprising result. The assumptions in question involve the agent's knowledge of elementary arithmetic, as well as a few further conditions such as closure of the knowledge base under modus ponens. Roughly, this result exploits Gödel-like self-referential sentences that contradict themselves.

What Montague showed is that any first-order theory that includes (any reasonably expressive theory of) arithmetic (such as Robinson arithmetic) and having as theorems $(Know\ \alpha) \rightarrow \alpha$ for each closed wff α, and also having as

the theorems *Know* α whenever α is a theorem, is inconsistent. (Here, and below, we suppress reifying quote marks on embedded wffs, writing *Know* α for *Know* "α", for ease of reading.)

Nevertheless, this may not be problematic in itself, since as noted earlier, knowledge may be a poor idea for AI. But (Thomason, 1980) found an equally surprising result for belief: he showed that under very similar assumptions, an agent's beliefs themselves will be inconsistent. Specifically, Thomason's result is this:

If an agent g believes (a suitable theory of) arithmetic and also g's beliefs (given as arguments to the predicate Bel) satisfy the following conditions:

1. $Bel\ \alpha \rightarrow Bel(Bel\ \alpha)$
2. $Bel(Bel\ \alpha \rightarrow \alpha)$
3. $Bel\ \alpha$ for all valid α
4. $Bel(\alpha \rightarrow \beta) \rightarrow (Bel\ \alpha \rightarrow Bel\ \beta)$

then g is inconsistent in the sense that g will believe all wffs.

(Perlis, 1988) has provided yet another related conundrum about belief: the ability to determine whether or not one has a given belief leads, under a few general and plausible conditions, to an inconsistent belief set. Again the technical device employed is that of a self-referential sentence. Thus even a perfectly omniscient reasoner would not be able to have perfect self-knowledge of its own beliefs. It is very puzzling why this should be so; it appears straightforward that one could build up such an ideal reasoner in stages by adding first one belief, then another, to achieve a fixed point. But, as Kripke emphasized in (Kripke, 1975), beliefs can refer to one another in very complex ways, and the truth of one belief thus can depend on others not even yet formed (as in "I will never believe anything you say").

These results are often taken as reason to abandon a "sentential" treatment of belief, where beliefs are taken to be formulas or sentences stored in a database, and in favor instead of a modal or propositional treatment in which beliefs are abstractions rather than concrete representations.

That is, the above results apply most directly to so-called syntactic treatments of belief (or knowledge), where the formal representation of a belief is that of a quoted wff or term to which a predicate symbol (such as a belief-predicate) may be applied. Modal treatments in which beliefs are construed as propositions do not quite so easily succumb to these problems.

Moreover, (des Rivières and Levesque, 1986) showed that one can smuggle the seeming virtues of modal logic into FOL, thereby vitiating the impact of the Montague and Thomason results. They removed the offending (inconsistent) sentences from Montague-Thomason settings, by allowing only those that were first-order "transcriptions" of modal wffs. (Morreau and Kraus, 1998) recently extended this work to its apparent logical extreme. On the other hand, it is hard to see why an agent should be restricted to using a weaker language than it is clearly capable of.

It turns out, however, that if we allow manipulation of either beliefs or *representations* of beliefs, in ways that are quite readily available to any routinely programmed computational agent, then the above negative (inconsistency) results apply equally to sentential and to modal treatments (Perlis, 1988).

A possible lesson that can be taken from this is that omniscient agents are not possible, either in practice (this is clear, since such agents require infinite storage for infinitely many beliefs) or in theory (due to the above limitations); and that instead of the sum total of all possible inferences an agent might make over its lifetime, we should focus on the actual inferences it makes in the short term to support its ongoing behaviors. This is the view taken in the active logic approach (Elgot-Drapkin and Perlis, 1990) in which the agent's belief base is at all times an (evolving) finite set of sentences which moreover can be inconsistent, and such an inconsistency, if discovered by the agent, is itself recorded in the belief base and used to prompt a corrective response by means of special inference rules. That is, an active logic can reason about its own ongoing inference process and take steps to influence its future behavior based on what it sees itself doing.

Another response to the above negative results is based on (Perlis, 1985; Perlis, 1988), namely to slightly weaken the contradictory assumptions. There are two ideas here: (i) to acknowledge that intuitions about formalisms that allow for self-referential expressions (this is where the arithmetic comes in) can lead to unintuitive consequences and thus we should restrict our usual intuitions to "safe" cases; and (ii) a particular way to distinguish safe cases can be derived from work of (Gilmore, 1974; Kripke, 1975). (i) is also the basis for the method of (des Rivières and Levesque, 1986) but that approach takes a more limited view of what is "safe" (transcriptions of modal wffs) than that afforded by (ii). The latter approach replaces key instances of α by the variant α^*. The starred version of α is in most cases of interest simply α itself. But when α contains, say, the predicate symbol Bel, preceded by negation, as in $\neg Bel\ \beta$, then its star is $(\neg Bel\ \beta)^* = Bel(\neg\beta)$. This simple "trick" (Gilmore, 1974) is sufficient to reinstate consistency to all the above axiomatizations, and includes all the cases of (des Rivières and Levesque, 1986) as well.

5.2 BELIEFS MUST BE INCONSISTENT

Although several approaches to the inconsistent foundational axiomatizations have been found, as indicated above, there are other matters for concern about inconsistency, of a different nature. There are reasons to suspect that commonsense reasoning is of its nature inconsistent (Perlis, 1996), having to do largely with the unreliability of incoming data in a complex world. For one simple case consider the following: an agent receives two pieces of data: (i) John is tall, and (ii) Mr. Smith is short. Both are from generally reliable sources S_1 and S_2, and so both (i) and (ii) are accepted as beliefs. But now the agent invokes its other beliefs, including that Mr. Smith and John are the same person, and a contradiction arises. This is not due to an error in reasoning, nor to a faulty knowledge-representation, but to a problem with incoming data. In

the present example either the incoming data contains an mistake (one of the sources was in error) or the meaning of one or more expressions was misinterpreted (e.g., the John in question is John Jones, not John Smith; or "tall" and "short" have quite different meanings to sources S_1 and S_2.

There seems to be no formal way to enforce consistency, then, short of allowing the agent in question to interact only with other agents whose own belief bases have first been "sanitized" for mutual consistency, clearly an impossible requirement in general.

These considerations are closely related to the topic of belief revision. The latter on the face of it deals with the question of what to do when clashes are found between beliefs: which beliefs are to be revised, and how? But actual studies have so far concentrated mainly on a special case, namely given that an incoming piece of data is definitely to be accepted (what was called "recency prejudice" in (Perlis, 1996)) which older conflicting beliefs are to be sacrificed? The underlying issue was raised early in the nonmonotonic reasoning literature (Reiter, 1980b; McDermott and Doyle, 1980) but then largely ignored until the publication of (Alchourron et al., 1985). Even this, however, retained the principle of recency prejudice, so that the challenge of adjudicating between competing beliefs without advance knowledge that certain ones are to be retained, went largely unexplored.[9]

At the same time, a rather different approach to commonsense reasoning was being developed, that emphasized the evolving nature of reasoning; this was initially called "step logic" (Elgot-Drapkin and Perlis, 1990) and later "active logic" (Miller and Perlis, 1993; Gurney et al., 1997; Perlis et al., 1998; Traum and Andersen, 1999). In this approach, beliefs are viewed as coming and going as reasoning and action take place, and new beliefs must compete along with the old on equal footing, unless special circumstances dictate otherwise. Moreover, the encountering of a direct contradiction by an active logic system, say P and not-P, is dealt with by the same underlying process of evolving inference as any other reasoning; that is, there is no halt to the "normal" process of reasoning while a separate "revision" process recreates a totally consistent belief set. Instead, when a new belief not-P is formed that may directly contradict an older belief P, both are gradually assessed, and as long as the direct contradiction remains in the belief base, neither P nor not-P is used as the basis for further inference; and at the same time ordinary reasoning proceeds. Much remains to be done, but there is some hope that such a gradualist approach to inconsistency may be successful in various domains (Elgot-Drapkin et al., 1993).

5.3 ABSTRACT CONTENT

Let us return to Susie Q. If her friend Freddy usually goes to the gym at the same time as Susie, and if we tell her the gym is closed today, we can predict

[9]The issue is most salient when the incoming information is not immediately seen to be in conflict with existing beliefs, so that it can be accepted on a "generosity" default: information tends to be correct unless counter-evidence is at hand. But further thought may reveal a hidden conflict.

that she will try to tell Freddy. But what will she tell him? This we cannot predict in detail, only that she will somehow try to convey to him *that he should not try to use the gym today.* But she may do this in countless ways, including countless linguistic ways, e.g.:

- The gym is closed today
- Don't go to the gym today
- Do not try to use the gym today
- You can't use the gym today
- The gym is not open today
- Don't count on the gym today
- You'll have to go somewhere else to exercise today

What these have in common is her intention to warn him about the gym not being open, and their abstract shared content to that effect. It is that abstract content that, somehow, she wants to convey to him, and that we can predict she will so try. Moreover, she wants him to acquire that content as a belief so he can act on it too. It is not just that she wants him to avoid the gym; she wants him to have the actual belief about it so he can act upon it in as general a way as she. For instance, he can then tell other friends to avoid the gym even if he had no plans to go there himself.

How can we design a system to reason with abstract content? Or is it enough for it to reason with concrete syntactic representations (sentence-like entities) in a flexible way? And how flexible should that be? We would not be highly impressed with a robot that could tell Freddy "avoid the gym" but that would not refrain from telling him that even if another agent had just told him out loud seconds earlier "Don't go to the gym today." Our robot should understand that Freddy already has been informed of this and that "avoid the gym" would just be a useless repetition. That is, the ability to suitably relate sentences by similarity of content is part of the intelligent use of beliefs.

But we have no good theory of content at present. (Kripke, 1979) has in fact pointed out a specific puzzle about the content of beliefs. I will paraphrase one of his examples as follows: suppose Susie hears that the gym is closed today, but also hears her friend Freddy say "I'm going to play tennis at the club today", without her being aware that the club Freddy has in mind is in fact the gym where they both usually exercise. So she does not tell him the gym is closed. Now the puzzle is this: she seems to have come to hold two distinct beliefs, one with the content that a certain facility is open, and the other that it is not open; that is, she seems to have two contradictory beliefs about the same place: that it is open and that it is not open. If one adopts the "unique names hypothesis" (Reiter, 1980a) – that distinct names refer to distinct entities – then such a situation appears to be ruled out; however, there are various reasons not to make such an hypothesis: (i) it is highly implausible for a real agent, as this example shows; (ii) not only names but *descriptions*

are involved; and (iii) multiple agents cannot be assumed all to use the same descriptions – or names – for objects.

Note that Susie herself may be totally unaware of the content-contradiction lurking in her beliefs, even if she knows all logical consequences of her beliefs. The contradiction depends on semantics, or at least on an outside view of her syntax, and not simply that syntax itself.

It seems there is some confusion here in the notion of what it is for a belief to be "about" something: one belief is about the facility seen-as-gym and the other about the same facility seen-as-club. Yet no one has been able to explicate the idea of content, so important for commonsense reasoning and behavior, in a way that avoids the apparent contradiction. One possible solution-route is to accept the contradiction in content and show how this can be fit into a theory of the ascription of belief in a way that retains predictive power concerning behavior; here "ascription" of a belief E to an agent G is simply the assertion that G believes E. Thus one might ascribe contradictory beliefs to Susie: she believes of that facility (the one that happens to be both the club and the gym) that it is open and that it is closed.[10]

Another solution-route might be to unpack "aboutness" in such a way that teases apart the two senses (gym vs club) by which the facility is known. This is related to the distinction between *de re* and *de dicto* belief ascriptions. The former corresponds to semantics (what object is referred to), the latter to syntax: we say Susie believes *de re* that a particular facility itself is open/closed; and we say she believes *de dicto* that whatever is referred to by her syntax (e.g., "the gym") is, say, closed. However, there is considerable doubt whether belief *de re* – independent of a syntactic characterization – is a meaningful notion at all; see (Devitt, 1984).

6 CONCLUSION

Beliefs are complex and poorly understood phenomena, yet they seem to be the best hope we have for understanding intelligent behavior. While it may be that an autonomous intelligence will be designed and built without our understanding how it works, it seems a good bet that a deep understanding of beliefs (what they are, how they arise, how they affect behavior) could only give us a leg up on the design problem.

The main research issues I see are:

1. characterize beliefs more carefully in terms of agent behavior
2. integrate this characterization into powerful modes of default reasoning and belief revision
3. keep the baby (effective commonsense reasoning) and throw out the bathwater (consistency).

[10] Kripke urges that this is a problem of how to *ascribe* beliefs, rather than about belief simpliciter. In either case it is a problem for the designer of AI systems: how are we to understand or analyze agent beliefs if we cannot define *or* ascribe them with clarity?

References

Alchourron, C., G
change. *J. Sym*
Bacchus, F., Gr
foundations for
Churchland, P. (19
des Rivières, J.
treatments of
Aspects of Reas
Devitt, M. (1984)
T. A., and W
Volume IX: Ca
Minneapolis, M
Elgot-Drapkin, J.
Active logics:
Report UMIAC
and CSD.
Elgot-Drapkin, J.
concepts. *Jour*
2(1):75–98.
Fagin, R., Halper
Knowledge. MI
Gettier, E. (1963).
Gilmore, P. (1974)
In Jech, T., edi
Ginsberg, M., edi
Kaufmann.
Gurney, J., Perlis
using active log
13(3):391–413.
Hintikka, J. (1962
NY.
Konolige, K. (1986
Kripke, S. (1975).
716.
Kripke, S. A. (197
and Use: Paper
pages 239–283.
Levesque, H. (1984
National Confe
American Assoc
McDermott, D.
Intelligence, 13(
Miller, M. and P
action. In *Proce*
Society, pages 7

References

Alchourron, C., Gardenfors, P., and Makinson, D. (1985). On the logic of theory change. *J. Symbolic Logic*, 50:510–530.

Bacchus, F., Grove, A., Halpern, J., and Koller, D. (1993). Statistical foundations for default reasoning. In *IJCAI*, pages 563–569.

Churchland, P. (1984). *Matter and Consciousness.* MIT Press, Cambridge, MA.

des Rivières, J. and Levesque, H. (1986). The consistency of syntactical treatments of knowledge. In *Proceedings of the conference on Theoretical Aspects of Reasoning about Knowledge*, pages 115–130.

Devitt, M. (1984). Thoughts and their ascription. In French, P. A., Uehling, T. A., and Wettstein, H. K., editors, *Midwest Studies in Philosophy, Volume IX: Causation and Causal Theories.* University of Minnesota Press, Minneapolis, MN.

Elgot-Drapkin, J., Kraus, S., Miller, M., Nirkhe, M., and Perlis, D. (1993). Active logics: A unified formal approach to episodic reasoning. Technical Report UMIACS TR # 99-65, CS-TR # 4072, Univ of Maryland, UMIACS and CSD.

Elgot-Drapkin, J. and Perlis, D. (1990). Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2(1):75–98.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge.* MIT Press.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, 23:121–123.

Gilmore, P. (1974). The consistency of partial set theory without extensionality. In Jech, T., editor, *Axiomatic Set Theory*, pages 147–153. Amer.Math. Soc.

Ginsberg, M., editor (1987). *Readings in Nonmonotonic Reasoning.* Morgan Kaufmann.

Gurney, J., Perlis, D., and Purang, K. (1997). Interpreting presuppositions using active logic: From contexts to utterances. *Computational Intelligence*, 13(3):391–413.

Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press, Ithaca, NY.

Konolige, K. (1986). *A Deduction Model of Belief.* Pitman, London.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72:690–716.

Kripke, S. A. (1979). A puzzle about belief. In Margalit, A., editor, *Meaning and Use: Papers Presented at the Second Jerusalem Philosophical Encounter*, pages 239–283. D. Reidel.

Levesque, H. (1984). A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202, Austin, TX. American Association for Artificial Intelligence.

McDermott, D. and Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence*, 13(1,2):41–72.

Miller, M. and Perlis, D. (1993). Presentations and this and that: logic in action. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pages 747–752, Boulder, Colorado.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability. In *Modal and Many-Valued Logics (Acta Philosophica Fennica, vol. 16)*. Academic Bookstore, Helsinki. Reprinted in R. Montague (1974). *Formal Philosophy*, New Haven, pp. 286-302.

Morreau, M. and Kraus, S. (1998). Syntactical treatments of propositional attitudes. *Artificial Intelligence*, 106:161–177.

Nilsson, N. J. (1983). Artificial intelligence prepares for 2001. *AI Magazine*, 4(4):7–14.

Perlis, D. (1985). Languages with self reference I: Foundations. *Artificial Intelligence*, 25:301–322.

Perlis, D. (1986). On the consistency of commonsense reasoning. *Computational Intelligence*, 2:180–190.

Perlis, D. (1988). Languages with self reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34:179–212.

Perlis, D. (1996). Nine sources of inconsistency in commonsense reasoning. In *1996 Workshop on Commonsense Reasoning*, Stanford.

Perlis, D., Purang, K., and Andersen, C. (1998). Conversational adequacy: Mistakes are the essence. *International Journal of Human Computer Studies*, pages 553–575.

Reiter, R. (1980a). Equality and domain closure in first-order databases. *Journal of the ACM*, 27:235–249.

Reiter, R. (1980b). A logic for default reasoning. *Artificial Intelligence*, 13(1,2):81–132.

Thomason, R. (1980). A note on syntactical treatments of modality. *Synthese*, 44:391–395.

Traum, D. and Andersen, C. (1999). Representations of dialogue state for domain and task independent meta-dialogue. In *Proceedings of the IJCAI99 workshop: Knowledge And Reasoning in Practical Dialogue Systems*, pages 113–120.