Northwestern University

The Institute for the Learning Sciences

RIGOR MORTIS: A RESPONSE TO NILSSON'S 'LOGIC AND ARTIFICIAL INTELLIGENCE'

Technical Report # 17 • August, 1991

Lawrence Birnbaum

Established in 1989 with the support of The Arthur Andersen Worldwide Organization

RIGOR MORTIS: A RESPONSE TO NILSSON'S 'LOGIC AND ARTIFICIAL INTELLIGENCE'

Lawrence Birnbaum

August, 1991

The Institute for the Learning Sciences
Northwestern University
Evanston, IL 60201

This research was supported in part by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N00014-85-K-0108 and by the Air Force Office of Scientific Research under contract F49620-88-C-0058. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting, part of The Arthur Andersen Worldwide Organization. The Institute receives additional support from Ameritech, an Institute Partner, and from IBM.

Abstract

Logicism has contributed greatly to progress in AI by emphasizing the central role of mental content and representational vocabulary in intelligent systems. Unfortunately, the logicists' dream of a completely use-independent characterization of knowledge has drawn their attention away from these fundamental AI problems, leading instead to a concentration on purely formalistic issues in deductive inference and model-theoretic "semantics". In addition, their failure to resist the lure of formalistic modes of expression has unnecessarily curtailed the prospects for intellectual interaction with other AI researchers.

1 Introduction

A good friend recently told me about a discussion he had with one of his colleagues about what to teach in a one semester introductory artificial intelligence course for graduate students, given the rather limited time available in such a course. He proposed what he assumed were generally agreed to be central issues and techniques in AI—credit assignment in learning, means-ends analysis in problem-solving, the representation and use of abstract planning knowledge in the form of critics, and so on. Somewhat to his surprise, in his colleague's view one of the most important things for budding AI scientists to learn was... *Herbrand's theorem.*

I recount this anecdote here for two reasons. First, I suspect that most of us would have been surprised, as my friend was, by this response, and by the rather puzzling set of scientific and educational priorities that it reveals. It would, of course, be unfair to conclude that the scientific world-view displayed in this anecdote is representative of the logicist position in AI, at least as that position is portrayed by Nils Nilsson. Nevertheless, it is worth bearing in mind that, as the anecdote makes clear, this debate is not only—perhaps not even primarily—a technical one, but rather a question of scientific priorities.

The second reason I recount this anecdote here is to illustrate that when it comes to making a point, a good story is almost always more useful than a lot of abstract argumentation. We often lean more heavily on personal experience and specific stories than on "book learning" or other abstract knowledge, and we often draw general conclusions from a single experience, even when we know we shouldn't. Human reasoning is powerfully affected by concrete images, by illustrative anecdotes, and by memorable experiences. Of course, artificial intelligence is not psychology: Our goal is not to mimic human thought and behavior in minute detail. Nevertheless, such a pervasive characteristic of human thought cannot simply be ignored as if it were *prima facie* irrelevant to our own work. Although I can't imagine that anyone in AI seriously disagrees with the proposition that there are probably sound functional reasons why human thinking is the way it is, the point nevertheless often seems to need repeating. The role of concrete cases in reasoning is something that many of us think is an important piece of the puzzle of both artificial and human intelligence; it is also a good example of the kind of question that logicists never seem to address.

Of course, it is not necessarily fatal to the logicist enterprise that it addresses only a portion of the problems involved in artificial intelligence: Who would have expected otherwise? The answer, I'm afraid, is the logicists themselves. Despite Nilsson's rather sensible observation that "successful AI systems of the future will probably draw on a combination

of techniques...," logicists do not seem to view logicism as just one approach among many: They see it as the universal framework in terms of which everything else in AI must ultimately be expressed. For example, in his response to McDermott's (1987) "A critique of pure reason," Nilsson (1987) asserts that "While all AI researchers would acknowledge the general importance of procedures and procedural knowledge (as distinguished from declarative knowledge), they would seem to have no grounds for a special claim on those topics as a *subset* of computer science." In other words, in Nilsson's view AI is to be distinguished as a sub-area of computer science in general not by the *problems* it investigates—language understanding, learning, vision, planning, and so on—but by the *methods* it uses, and in fact by one and only one aspect of those methods, the use of declarative representations. Since Nilsson further makes it clear that in his view the use of declarative representations must, ultimately, entail embracing all of the apparatus of logic, the implication of this assertion is fairly obvious: Anything that doesn't fit the logicist paradigm—visual perception, goals, memory and indexing, attention, emotions, the control of physical movement, and so on—may be very nice computer science, thank you, but it isn't AI. This is not, it should be clear, a scientific proposition, but rather a political one, and as such its role in our field deserves careful scrutiny.

In addition to these sorts of political concerns, however, there are also good technical reasons for doubting the utility—or at the very least, the *special* utility—of logic, strictly speaking, as a framework for knowledge representation. I say "strictly speaking" because, in a broad sense, any computational scheme for knowledge representation and reasoning could be considered some form of logic, even connectionism. Moreover, most work in AI shares the logicist commitment to the centrality of explicit symbolic representations in mental processes.

How then does logicism differ from other approaches to AI? In my view, it is distinguished by two additional commitments. The first is its emphasis on sound, deductive inference, in the belief—for the most part implicit—that such inference plays a privileged role in mental life (see McDermott, 1987). As a result of this emphasis, logicists tend to ignore other sorts of reasoning that seem quite central to intelligent behavior—probabilistic reasoning, reasoning from examples or by analogy, and reasoning based on the formation of faulty but useful conjectures and their subsequent elaboration and debugging, to name a few—or else, attempt (unsuccessfully, in my view) to re-cast plausible inference of the simplest sort as some species of sound inference.

The second distinguishing feature of logicism is the presumption that model-theoretic "semantics" is somehow central to knowledge representation in AI.[1] The primary justifica-

[1] I place scare quotes around "semantics" in this sense to emphasize that in logic this is a technical term,

tion for this presumption is its putative explanatory role, namely, that it is necessary in order to correctly characterize what it means for an agent to know or believe something, and thus to specify precisely what a given agent knows or believes. This, I take it, is Nilsson's argument.

Towards this claim, non-logicists seem to be of two minds: Some disagree with it—and I will explain why later in this paper—while others just don't see what difference it makes. In the face of the latter reaction—indifference—logicists often take another tack, and attempt to justify their preoccupation with model-theoretic "semantics" on methodological, rather than explanatory, grounds. In this vein, they stress its utility as a heuristic aid in solving knowledge representation problems, or to enable the AI researcher to prove things about his program, and so on. Whether the things that can be so proven are of any particular interest is debatable (see DeMillo, Lipton, and Perlis, 1979, for some arguments as to why proving that programs meet certain formal specifications is unlikely to be either possible or useful in software engineering; it seems doubtful that AI systems will prove more tractable in this regard). Whether such a "semantics" is in fact heuristically useful in solving representation problems is more difficult to debate, since that is a matter of personal taste. People should, obviously, work in the manner that they find most congenial and productive.

2 The good news

The above criticisms notwithstanding—and I will take them up in greater detail shortly—it cannot be denied that logic and logicists have contributed a great deal to AI. Perhaps the logicists' most important contribution has been to focus attention on the fact that it is the *content* of our knowledge, and the concepts in terms of which that knowledge is expressed—what logicists refer to as "ontological" issues—that lie at the heart of our ability to think. Their arguments have played a key role in legitimizing the study of representations from the perspective of how well they capture certain contents, independently of the details of the particular processes that might employ them. Hayes's (1979) "Naive physics manifesto," in particular, is a persuasive and historically important defense of this idea. As Nilsson puts it, "The most important part of 'the AI problem' involves inventing an appropriate conceptualization..." In fairness, however, the credit for this insight cannot be assigned solely to the logicists: Very much the same point can and has been made without any special commitment to logic, for example by Feigenbaum (1977), Schank and Abelson (1977), and

and that the theory to which it refers may or may not turn out to be a correct account of the meaning of mental representations.

Newell (1982), among others. Moreover, as Nilsson acknowledges, "the [logicist] approach to AI carries with it no special insights into what conceptualizations to use."

The technical apparatus of logic itself—a clear and unambiguous syntax, the "ability to formulate disjunctions, negations, and universally and existentially quantified" expressions—indisputably plays an important role in AI, as does the technology of mechanical theorem proving. Although I am sympathetic with the view that it is a mistake to attempt to embed "scruffy" thinking in "neat" systems[2]—and that the real question is how "neat" thinking can emerge from a "scruffy" system—expressive apparatus of the sort provided by logic seems indispensable, especially when it comes to representing abstract concepts.[3] On the other hand, Nilsson's implicit criticism of knowledge representation schemes that lack this technical apparatus probably misses the point. Much research that might be criticized on these grounds has simply been directed towards other issues, primarily issues of content and conceptual vocabulary, or of memory organization and efficient access for a particular set of tasks.

Finally, there can be no question that logicists have led the battle for declarative representation in AI, a battle that they have largely won—though again, not entirely single-handedly. A particularly compelling argument, attributed by Nilsson to McCarthy, is that declaratively represented knowledge "[can] be used by the machine even for purposes unforeseen by the machine's designer..." Putting this in somewhat different terms, there can be no question that cross-domain and -purpose application of knowledge is an important functional constraint on representations, and that declarative representations seem to meet this requirement better than anything else we know of. However, putting it this way makes it clear that what is at stake here is a type of *learning*—the ability to apply knowledge acquired in one situation, for one purpose, to other, different situations and purposes.

Yet, rather oddly, Nilsson, and logicists generally, pay no attention to this issue. There are several reasons for this, but the main one seems to be their attachment to sound inference and model-theoretic "semantics". For example, in his discussion of inference, Nilsson argues that

> Often, ... the new sentence constructed from ones already in memory does not tell us anything new about the world. All of the models of [the sentences already in memory] are also models of [the new sentence]. Thus, adding [the

[2]See Abelson (1981) for an enlightening discussion of these terms.

[3]Which aren't necessarily very technical or abstruse: Try representing the concept of "helping" in propositional logic.

> new sentence] to [memory] does not reduce the set of models. What the new sentence tells us was already implicitly said by the sentences from which it was constructed.

Indeed, logicists sometimes go so far as to assert that sound inference cannot, in principle, generate any new knowledge. On this account of what knowledge is, or of what makes it "new", the problem of applying lessons learned in one domain for one purpose to other domains and other purposes doesn't exist, because it is already solved. Unfortunately, if this isn't true—and it strikes me as a rather dubious proposition on which to bet a research program—then we must conclude that the problem cannot even be properly *characterized* within the logicist framework. However, if the problem of cross-domain and -purpose application of knowledge cannot even be *characterized* within the logicist framework, then we have good reason to doubt that logic, as construed within that framework, in fact appropriately addresses the functional issues raised by the problem. Since, as Nilsson himself argues, this problem provides the fundamental justification for declarative representations in the first place, the logicists have some explaining to do.

3 The bad news

What drives logicists to adopt the obviously unrealistic position that inference does not change what an agent knows? It is their devotion to model-theoretic "semantics". And what motivates this devotion? Nilsson puts it as follows:

> Those designers who would claim that their machines possess "knowledge" about the world are obliged to say something about what that claim means. The fact that a machine's knowledge base has an expression in it like [(forall (x) (if (box x) (green x)))], for example, doesn't by itself justify the claim that the machine *believes* all boxes are green.

This is uncontroversial, as far as it goes. The leap of faith in the logicist program is the presumption that, in saying something about what it means for a machine to have beliefs, AI is obliged to reiterate a theory of how logical symbols are to be interpreted, developed over the last century in logic and mathematics for fundamentally different purposes—in particular, for proving the soundness and completeness of inference methods. Nilsson offers no argument why this should be so. But McDermott (1987) is a bit more open about how

the logicists arrived at this position: "The notation we use ... must have a semantics; so it must have a Tarskian semantics, because there is no other candidate"—or to put this another way: *"You have a* better *theory?"*

I must admit that I do not have a better theory—at least, not one that would satisfy the logicists. But the absence of an alternative theory does not make a bad theory good. I find it difficult to understand the zeal with which logicists embrace and defend a theory that has so many problematic implications. Trying to define "knowledge" and "belief" at our current stage in theorizing about the mind is like biologists trying to define "life" a hundred years ago. Rather than seeing this as a complicated puzzle to be resolved by artificial intelligence and other cognitive sciences as they progress, logicists assume that the question has a simple, definitive answer, that logic has provided this definitive answer, and that all AI has to do is work out the details.

The obvious alternative to a model-theoretic "semantics" is a *functional* semantics, based on the idea that representations get their meaning by virtue of their causal role in the mental processes of the organism, and ultimately, in perception and action. On such a view, the meaning of a term is not tied to the inferences that it could *in principle* license, but to those that it actually licenses *in practice.* The concept "prime number" does not mean the same thing to me as it does to a number theorist; and its meaning for me would change if I studied some number theory.

The problem with such an approach, from a logicist perspective, is that a theory of meaning based on functional role doesn't allow for a specification of what an organism knows independently of what it does. This is exactly right. Moreover, such a situation does not preclude applying the same knowledge in different domains, and for different purposes. Indeed, it seems clear that Nilsson's chief argument for the logicist position is based on a false dichotomy: Just because knowledge cannot be specified in *complete* independence of use doesn't mean that it can't be specified in a way that enables its application to many *different* uses. But for logicists, it isn't sufficient that the ability to apply the same knowledge in different domains, and for different purposes, be a functional constraint on mental representations—something that is, all other things being equal, to be desired. Such a formulation implies that generality of this sort might be involved in "grubby" engineering trade-offs (to use McDermott's colorful description), and this is exactly what logicism is trying to escape in the first place. As Dennett (1988) cannily observes, logicism is the chief representative, within AI, of a belief in the "dignity and purity of the Crystalline Mind," and of the concomitant notion that psychology must be more like physics than like engineering or biology.

The upshot is that logicists believe there is no alternative to embracing model-theoretic "semantics" for mental representations. The major stumbling block in any straightforward application of this approach is, as has often been noted, consistency—or more precisely, the lack of it (see, e.g., Minsky, 1974; Hewitt, 1987).[4] If the beliefs of an organism are inconsistent, then there is no model of those beliefs. This is problematic for a number of reasons. First, it means that the content of the organism's knowledge, which Nilsson asserts should be characterized as the "intended" model of the sentences representing that knowledge, cannot in fact be so characterized: In the technical sense, there are no such models. It follows that whatever the relationship between the organism's representations and the content they express, that relationship cannot be described by a model-theoretic "semantics". The second problem, of course, is that if the inference processes of the organism are to be construed as some form of logical deduction, then if its beliefs are inconsistent, anything at all can be deduced. Because its beliefs have no model, all models of its beliefs are also models of any other belief it might entertain.

Inconsistency poses such a severe problem for model-theoretic "semantics" because of its extremely "holistic" nature, in that the meaning of an individual symbol in a logical theory is determined by the set of models consistent with the entire theory in which it appears. As Hayes (1979) puts it, "a token [in a formal theory] means a concept if, in every possible model of *the formalisation taken as a whole,* that token denotes an entity which one would agree was a satisfactory instantiation of the concept in the possible state of affairs represented by the model." There is a certain intuitive appeal to this holism; indeed, in any functional approach to semantics, the meaning of a symbol similarly depends upon the entire system in which it is embedded. The problem is that model-theoretic "semantics" takes the holism of meaning to its extreme: Either a theory is completely consistent, or it has no models, and hence no way to determine meaning at all. A knowledge base with a single bug that makes it inconsistent is as meaningless and incoherent as gibberish. There is no graded notion of coherence in this conception of semantics, and it seems clear that one is needed.

A contributing factor here is the logicist assumption, generally implicit, that a successful organism's beliefs about the world are such that the real world is in fact a model of those beliefs. As Nilsson puts it, "the designer attempts to specify a set of conceptualizations such that, whatever the world is, he guesses it is a member of the set." In the talk upon which his paper is based, he went somewhat further and asserted that "the conceptualization *is* the world! (To the extent that it isn't and matters that it isn't—change the conceptualization.)" This assumption insulates logicists from one of the two potential sources of inconsistency,

[4]See Lakoff (1987) and Woods (1987) for discussions of some other difficulties.

namely, the possibility that the set of beliefs about the world that an organism brings to any particular problem situation, whether innate or acquired, are themselves inherently inconsistent.[5] The ploy succeeds because we all assume that the real world is consistent; so if the real world is a model of the organism's beliefs, they must be consistent too. In his paper, Nilsson backs off to a certain extent from the claim that an organism's beliefs will have the real world as a model. Nevertheless, he asserts, without argument, that this is what we should aspire to. The fact of the matter is, however, that we have no reason to believe this is so. Naive physics is not physics, it is *psychology*: An organism's conceptualization of the world differs from the world in fundamental ways, and for very good functional reasons. If fidelity to the real world really were the paramount constraint on conceptualizations, then AI would seem best served simply by axiomatizing the latest theories of the physicists. If logicists don't believe this, then they must accept the fact that the primary constraints on conceptualizations are *pragmatic,* derived from the need to perform effectively in real-world tasks—in other words, that the content of our beliefs is determined in great measure by the uses which they must serve.

Now, is it reasonable to expect organisms to have perfectly consistent beliefs? Or is it more reasonable to expect that they will have some conflicting beliefs? Even the most committed logicists seem to acknowledge that the latter is more likely. Since an organism's beliefs arise, ultimately, from perception and learning, any mistakes in perception or learning would give rise to erroneous beliefs, and these would be likely to conflict with true beliefs of the organism, if not immediately then eventually—indeed, they had better, if the organism is ever to discover such errors.

One place to uncover concrete examples of contradictory beliefs is to consider questions about which we feel ambivalent or uncertain, such as tough moral questions. For example, I believe, on balance, that abortions ought to be freely available. On the other hand, I believe that killing a human being is unacceptable except in self-defense. So in order to reconcile these two beliefs, I have decided that fetuses are not human beings. Nevertheless, I also believe that abortions are somehow a bad thing, and should be avoided if possible. Moreover, I recently learned that there are reasons to believe that fetal tissue will be particularly effective in transplants, e.g., to cure Parkinson's disease. Should this prove to be the case, fetal tissue may come to be in high demand. This raises the following question: Would it be moral for a woman to conceive for the sole purpose of having an abortion to provide such tissue? I have qualms about this. Now the problem is, given my ostensible belief that fetuses are not human beings, I'm not quite sure why I have any of these reservations. The best explanation I can give for my ambivalence about this issue is

[5]The other possible source of inconsistency is unsound inference.

that, in fact, several conflicting beliefs bear on it.

But just as the person who is led down the garden path in an argument will squirm and wiggle and look for an implicit assumption that will let him off the hook, so the logicist argues that there are only *apparent* contradictions in an organism's beliefs: In fact, there are always additional qualifications attached to one or the other of two apparently inconsistent beliefs, and the organism has simply assumed, mistakenly, that all of these additional conditions on its beliefs hold true. Moreover, given any putative example of conflicting beliefs, this trick can always be pulled, and the argument can be made that the beliefs in question are in fact so qualified as to eliminate the apparent contradiction. This leads us, then, to the logicist characterization of plausible inference as deduction, given some extra assumptions that may or may not turn out to be true—in other words, non-monotonic logic.

In many ways, this is an attractive vision. The problem with this vision is that it seems difficult to characterize ahead of time all of the extra assumptions that one is in fact committed to in drawing a given conclusion, if one wants to view the drawing of that conclusion as a form of deduction. Nilsson makes this point using McCarthy's example of what might turn out to be involved in trying to determine the conditions under which we can infer that a car will start. The upshot is a kind of stand-off: Given any particular example of conflicting beliefs, logicists can plausibly argue that the beliefs in question are implicitly qualified so as not to conflict. But we are left with the suspicion that given more examples, such qualifications would have to be extended indefinitely, to the point of the ridiculous.

In any case, Nilsson concedes that organisms will have inconsistent beliefs in his discussion of the "reification" of theories. He accepts the view (espoused by Hewitt, 1987, among others) that problem-solving depends on the manipulation of relatively fragmented and mutually inconsistent *microtheories*—each perhaps internally consistent, and each constituting a valid way of looking at a problem:

> We might reify whole theories. This will allow us to say, for example, that some [set of beliefs] is more appropriate than some [other set of beliefs] when confronted with problems of diagnosing bacterial infections. Scientists are used to having different—even contradictory—theories to explain reality... Each is useful in certain circumstances.

I agree that it is useful to have contradictory microtheories. But I find it difficult to understand how Nilsson reconciles this belief with the logicist program. What the phrase

"each [theory] is useful in different circumstances" *really* means is that each is useful for different *purposes.* Such a proposal seems utterly inconsistent with the logicist dream of specifying knowledge in complete independence of use. Moreover, what status does Nilsson assign to the elements of the inconsistent theories that scientists "have": Are they beliefs, or not?[6] If so, then what is their model? For if model-theoretic "semantics" actually provides a correct account of what it means for an agent to know or believe something, as Nilsson asserts, then the elements of these inconsistent theories must have a model. But by virtue of their inconsistency, this is impossible. Alternatively, I suppose, Nilsson could argue that the elements of such inconsistent theories are not in fact beliefs. In that case, we may freely admit that logicism has provided a satisfactory account of what it means for organisms to have beliefs; it just turns out that beliefs, so construed, play little or no role in their reasoning processes.

The hard-core logicist response to this dilemma has been enunciated, if I understand him correctly, by McCarthy. Seemingly inconsistent microtheories, taken together, *do* have models: We must simply qualify *every* belief in *every* microtheory, if necessary, with the condition that the objects it concerns are not abnormal, conjoin all of the resulting microtheories, and then use *circumscription* (McCarthy, 1980) to limit the models of the resulting qualified and unified theory to those in which the number of abnormal objects is as small as possible.

Despite all its technical bravado, however, this proposal strikes me as a desperate strategem. *For there is no guarantee that the resulting models will bear any resemblance to the intended models underlying the initial (unqualified) microtheories.* It follows that the inferences we will be entitled to draw after following this procedure will, almost surely, be different from those we had in mind when the microtheories were originally constructed. Indeed, on the view that the meaning of a term is tied to the inferences it helps to license, the meanings of the concepts involved in a given microtheory are likely to be considerably different from what we originally intended.

What is most fatal to this proposal, however, is that we will almost surely be unable to tell whether or not these sorts of divergences from our original intentions have actually arisen in any given case. For although McCarthy's strategem guarantees that the resulting unified theory has models, not only does it fail to guarantee that those models have the properties we need, *it doesn't even guarantee that we know what those models are.* I take one of the larger lessons of Hanks' and McDermott's (1987) "shooting problem" to be that even given a small, simple set of initial beliefs, it is quite difficult to determine the models

[6] I am indebted to Drew McDermott for pointing this problem out to me.

of those beliefs permitted by circumscription.[7] Since the task of determining the models permitted by circumscription has proven so difficult for a small set of three or four beliefs, it seems unimaginable that it will be possible to determine those permitted a large knowledge base of conjoined, qualified microtheories. In short, although McCarthy's strategem does make it possible to ensure the consistency of a set of beliefs, and the existence of models for that set, the cost it exacts is that we no longer know what in fact those beliefs say, or whether the inferences we need actually follow from them. His proposal destroys semantics, in any meaningful sense, in order to save "semantics" in a technical sense.

Of course, the deeper question here is whether our conceptualizations of the world are or can be consistent and independent of use: The technical difficulties that surround model-theoretic "semantics" are, for the most part, a consequence of attempting to pursue the logicist dream of a completely context-free characterization of knowledge too far. And as I have already pointed out, Nilsson himself asserts, in his discussion of "reification", that the answer to this deeper question is no: Conceptualizations are not independent of use. However, this is true in a way that is even stronger than he implies. Nilsson's point is that different problem situations and different goals will require different—and incompatible—conceptualizations of the world. The fact is, however, that even if they *are* compatible, all conceptualizations of a situation are not the same.

This point was first made by Cordell Green in his discussion of QA3 (Green, 1969), one of the earliest serious attempts to apply logical methods to problem solving. QA3 brought together the situation calculus (McCarthy, 1968) and resolution theorem proving (Robinson, 1965) for the first time, and applied them to planning and automatic programming. In his experiments with it, Green discovered something interesting: When applied to relatively complicated problems, such as the Tower of Hanoi, or writing a program for merge sort, whether or not QA3 could find a solution depended critically on how the axioms were formulated. The point he made then is still true now: One cannot in fact just "write down the axioms" in blissful ignorance of their intended uses. Logically equivalent ways of conceptualizing the world are not functionally equivalent. Nilsson seems to acknowledge this point when he writes that "Because the actions emitted by [the function that maps from an organism's beliefs to its actions] depend on the syntactic form of the sentences [representing those beliefs in the organism's memory], it is necessary ... to be able to rewrite these sentences in the form appropriate to the task at hand." But I think it is fair to say that

[7]To review briefly, the Hanks-McDermott problem is this: Given the event sequence (1) Fred is born, (2) A gun is loaded, and (3) Fred is shot with the gun, plus the belief that if someone is shot with a loaded gun they will die, infer that Fred is dead. Hanks and McDermott have shown that circumscription, along with other forms of non-monotonic reasoning, permit unintended models in which the gun becomes unloaded after (2) but before (3), and Fred remains alive.

this acknowledgment, by focussing on the trivial and obvious dependence of an organism's effectors on the form of the signals that control them, draws attention away from the larger question of how the differences between logically equivalent conceptualizations might affect the task of reasoning itself, and thus seems to imply that such differences are far less important to the invention of appropriate conceptualizations than they actually are.

There is yet a further lesson about the interdependence of ontology and function to be drawn from Green's work. In his analysis of QA3's weaknesses, he made the following point:

> Let us divide information [needed for automatic programming problems] into three types: (1) Information concerning the problem description and semantics... (2) Information concerning the target programming language... (3) Information concerning the interrelation of the problem and the target language... In the axiom systems presented, no distinction is made between such classes of information. Consequently, during the search for a proof the theorem prover might attempt to use axioms of type 1 for purposes where it needs information of type 2. Such attempts lead nowhere and generate useless clauses. However, ... we can place in the proof strategy our knowledge of when such information is to be used, thus leading to more efficient proofs. (Green, 1969, p. 235)

In other words, concepts about the world must be categorized in useful ways—in terms of abstractions whose definition is motivated not solely by the world itself, but by the need to organize knowledge about the world for effective problem solving. This is by now a widely accepted proposition. Still, I think it is fair to wonder about the status of such abstract categories within the logicist framework. As far as logic is concerned, an abstract category of this sort is really just a shorthand notation for the disjunction of all the concepts which it categorizes. If one were simply to eliminate all such abstractions from a set of axioms, and replace them with the equivalent disjunctions, the resulting set of axioms would be logically equivalent to the original set. Indeed, any set of such abstractions is as good as the next, logically speaking, since none makes the slightest difference to the conclusions that one can draw about the world. Thus, abstract concepts defined in order to usefully categorize knowledge are ontologically—and semantically—vacuous from a logicist perspective. In sum, if the "conceptualization" represented by a given knowledge base *is* (or should be) the world, as Nilsson asserts, then any additional concepts, categories, or relations defined in order to make reasoning more effective are not included by this term—it covers only a subset of the concepts, categories, and relations in terms of which we understand the world.

It is, however, with conceptualizations in this larger sense that AI must be concerned, as Minsky (1974) has forcefully argued.

It is worth pointing out that a similar argument forms a portion of Imre Lakatos's (1976) brilliant and entertaining critique of the logicist approach to mathematics, *Proofs and Refutations.* Among other things, Lakatos shows how utterly inadequate the logicist conceptions of definitions as "theoretically dispensable but typographically convenient abbreviatory devices"—and ultimately, of proofs as formal deductions—actually are to explain what our mathematical concepts are and how they develop. It need hardly be added that the doubts Lakatos raises about whether logic can adequately account for how we think about mathematics should dampen anyone's enthusiasm about its ability to adequately account for how we think about everything else.

Given all of these difficulties, both theoretical and empirical, why do logicists continue to adhere to the position that conceptualizations are independent of use, and to the concomitant notion of model-theoretic "semantics"? I think the reasons have more to do with methodological hopes (and fears) than anything else. Process models of intelligence depend on knowledge; if the knowledge can't be formulated independently of the process models, where do we begin? How can we write down what a program needs to know before we know what the program looks like? And how can we write the program if we don't have some theory of what it needs to know? The logicists see the claim that knowledge can be formulated entirely independently of use as the only way to avoid this vicious circle.

It seems to me that this apparent circularity is based on an overly simplistic view of science (and for that matter, programming). We might, instead, view the process of constructing AI theories as proceeding by successive approximations, starting with an approximate theory of the necessary knowledge, constructing a preliminary algorithm, and then refining them both in concert. In reality, of course, this is exactly what everybody does. Moreover, viewed within the context of this more realistic characterization of AI methodology, our preliminary and descriptive theories of the contents of mental structures *sans* process models are merely that: preliminary and descriptive. There is no need to become obsessed with the formal properties of the notations we use in constructing such theories, because there is no reason to believe that these will play any explanatory role in the final theory.

To put this another way, any attempt to specify the contents of mental representations in complete independence of use is probably doomed to failure. Nevertheless, it is worth *pretending* that this is not so, since important and useful investigations of the knowledge necessary in order to behave intelligently are likely to result, and indeed *have* resulted, from

such attempts. The question is how seriously we need to worry about the notations in which such investigations are carried out. As Hayes (1979) points out, "Initially, the formalisations need be little more than carefully-worded English sentences. One can make considerable progress on ontological issues, for example, without actually *formalising* anything." He then goes on to argue that, in short order, it will be necessary to express such intuitions formally. I agree with him, but it seems to me that such formalization must take place within the context of a set of actual tasks, so that we have some idea of the purposes to which the knowledge must actually be put. I don't mean to deny that the ability to apply the same knowledge to many different tasks should play a role in determining the appropriate formalization. But the only compelling reason to argue that such formalization must be in terms of logic, in the narrow sense, is if you assume that the knowledge will be applied by a deductive engine—i.e., a theorem prover. Without this assumption, the methodological imperative for formalization in terms of logic just isn't there. It is, after all, insights into "ontological issues" that are the point of the investigation in the first place. Painstaking attention to the formal properties of the notation in which such insights are expressed is misplaced. In those cases where content theories of what we know about some domain are expressed in English, for example, it seems difficult to imagine that Hayes or anybody else would advocate spending a lot of time worrying about linguistics.

This brings us, finally, to the logicists' preoccupation with deduction. The most straightforward argument for formalizing knowledge in logic, narrowly construed, is simply that if we do not do so, then we cannot rely on deduction, narrowly construed, as our model of reasoning. In my view, however, this has resulted in a reversal of priorities: The logicists have been led to embrace deduction as the process by which knowledge is applied in order to motivate the use of logic, rather than the other way around (McDermott, 1987, makes a similar point). Indeed, one can find clear evidence for this underlying motivation in the literature. Consider, for example, the following quote from Patel-Schneider (1985):

> [The undecidability of first order logic] has led to many attempts to create [knowledge representation] systems based on [first order logic] that always produce answers. Most of these systems retain the syntax of [first order logic] while modifying its inferences in some way. The crudest of them simply take a theorem prover for [first order logic] and place some *ad hoc* restrictions on it, such as terminating the search for a proof after a pre-set amount of time or a certain number of proof steps. Such modifications produce systems that cannot be given an adequate semantics and have no means of completely characterizing answers except by referring to the actions of the modified theorem prover. This destroys most of the advantages of using logic in the first place.

I agree with much of Patel-Schneider's discussion here: The sort of functionally motivated deviations from deductive inference which he stigmatizes as *"ad hoc"*[8] do, it seems to me, call into question the applicability and relevance of model-theoretic "semantics". But whereas this suggests to me that it is model-theoretic "semantics" that should be dispensed with, Patel-Schneider goes on to make it clear that in his view it is any deviation from pure deduction which must be abjured.[9] In this, the purest and most extreme form of the logicist world-view, the primary constraint on an inference process is not whether or how well it performs in some realistic task, but whether it can be given an "adequate" logical characterization. Needless to say, what we can expect from this approach are "logically characterized" systems that don't do anything particularly interesting. The literature is full of this sort of "result".

4 Residual problems

In this section, I would like to address some residual questions, having more to do with Nilsson's paper than logicism *per se.*

(1) Addlists-deletelists and the frame problem. In his discussion of the frame problem (McCarthy and Hayes, 1969)—the problem of determining what doesn't change in the world when an action is performed—Nilsson refers briefly to a number of approaches that have been suggested. Most of these fall squarely in the logicist camp, but at the head of the list he includes his and Fikes's use of addlists and deletelists in the STRIPS problem-solving system (Fikes and Nilsson, 1971). Leaving aside any perfectly understandable personal fondness that Nilsson might have for this approach, its inclusion on such a list, in a paper touting the virtues of logicism, seems completely incomprehensible. For whatever its other merits or deficiencies, this approach represents the antithesis of logicism. It may allow problem solving systems to infer what changes and what doesn't change when actions are performed in certain simple domains—domains in which, e.g., the results of actions are not conditional on the state in which they are executed—but those inferences are by no means "logical" in the strict sense demanded by logicism. For better or for worse, the approach

[8]See Schank (1978) for an explanation of the political role that this term plays in the cognitive sciences. What *"ad hoc"* really means, of course, is "for a special case." Attempts to deal with fundamental constraints such as undecidability hardly strike me as being for a special case.

[9]Nor is he alone in this view: Levesque (1986) dismisses such efforts as "pseudo-solutions" on the grounds that we cannot guarantee that the resulting reasoning systems will always get the right answer. I cannot imagine why anyone believes that we will be able to guarantee that intelligent systems will always get the right answer. Does Levesque believe that *people* always get the right answer?

simply employs a procedure that updates the data base by adding the assertions on the addlist, removing those on the deletelist, and leaving everything else alone. This may or may not be the right thing to do in certain circumstances, but surely it isn't the *logicist* thing to do—or if it is, then the label loses all significance.

(2) Reasoning with models. Nilsson argues that in many cases it is possible to construct an analog of the "intended model" of a logical theory, or at least of a portion of it, and to reason about what is true in the model by directly examining it—a process that might turn out to be considerably cheaper than theorem proving. As he puts it, "Because [sound logical deduction] guarantees that a derived sentence is satisfied by a whole set of interpretations (including the intended one), it may be too strong for most purposes—and thus too expensive. All that we really need is to know whether the *intended* interpretation is a model of the sentence in question." The question this raises is, why would we ever care about anything else? I also wonder what the logicist position on the semantics (in the informal sense) of such a model might be. I imagine that logicists would be tempted to argue that this is provided by the "semantics" of the logical theory of which it is the "intended model"—but that is of course circular. Indeed, it turns the relationship on its head. The only coherent option, then, is simply to examine the representation of the model and provide rigorous, informal arguments that it has the right properties. This, however, calls into question the methodological utility of model-theoretic "semantics" in general: Once again, the question is, if such arguments are good enough in this case, why would we ever want to do anything else?

(3) The practicality of deduction. Nilsson acknowledges that logical deduction is computationally expensive (that's a bit of an understatement), but argues that it is nevertheless practical in many cases. As evidence, he claims that "many large-scale AI systems depend heavily on predicate calculus representations and reasoning methods." He then goes on to list three such systems, among which is included Appelt's (1985) natural language generator. I admire Appelt's work a great deal. However, the program to which Nilsson refers cannot by any stretch of the imagination be called a "large-scale AI system". It is an experimental AI program that handles a few interesting examples. Moreover, Appelt has been reputed to assert, in a mock-boastful fashion, that his program is the slowest generator in existence. What he means by this, of course, is that his program implements the most detailed and faithful model of generation. The point remains, however, that this sort of work is hardly an argument for the large-scale practicability of logical methods.

5 A plea for plain speaking

Having reviewed what I see as the strong and weak points of the logicist enterprise, I would like to return to a theme I touched at the very beginning of this paper. It seems to me that the primary distinction between logicism and AI in general is not, as the logicists themselves seem to believe, a matter of technical issues, but rather a question of scientific world-view, of priorities and ways of looking at problems. Logicism represents, in my view, an understandable longing for a technical basis upon which to ground AI research. In this, it seems similar to certain forms of connectionism, a point to which Charniak (1988) alludes when he identifies logicism with a larger trend that he terms "mathism". Although the assumptions, research programs, and even the personalities of the adherents differ radically, nonetheless from a sociological perspective, connectionism and logicism share a great deal. Connectionism has its neural nets, its energy function equations, its convergence theorems; logicism has its axiom schemata, its model theoretic "semantics", its completeness theorems. Both appeal outside of AI for their foundations, logicism to analytic philosophy and mathematical logic, and connectionism to neurobiology and physics. Both represent something of a backlash against the dominance of expert systems in AI over the past ten to fifteen years.

The chief problem in both cases is that the appeal outside of the common heritage of AI inevitably makes communication more difficult, inhibiting the intellectual give and take that is so important to making progress on difficult issues. It takes a great deal of effort to read logicist papers, and unfortunately, the actual ideas and results being reported, once understood, rarely prove to be worth the cost in man hours required to put them in more straightforward terms. Nor am I alone in feeling this way. Forbus (1987) indicates a similar annoyance with the opacity of logicist formulations:

> Anyone can write axioms. The problem is figuring out what should be said, and saying it precisely. ... One can have ad hoc axiomatic theories just as easily as ad hoc theories stated in natural language. Everyone has their favorite examples. (I won't mention mine here, since it will only raise heat without shedding light.) The major difference is that, because more detail is involved, it usually takes more work to uncover bugs in axiomatic theories than in theories stated in English.

However, I disagree profoundly with Forbus's attribution of the problem to the allegedly greater detail to be found in logicist theories. I suspect, in fact, that he was merely being

polite. Certainly, the work that he and others have done in qualitative physical reasoning is far more detailed than anything that has been produced by the logicists.

The inevitable result of the logicists' private language is that they end up talking primarily among themselves, and the larger dialectic into which they might enter is short-circuited. This is a loss both to the logicists—who are missing out on useful ideas, comments, and criticisms that non-logicists might offer—and to the rest of us. The fact of the matter is that one does not need a detailed understanding of circumscription, for example, to have useful and right-headed ideas about plausible reasoning. Unfortunately, the social utility of pretending otherwise has proven too great a temptation to logicists. Consider, for example, Nilsson's discussion outlining the motivations underlying the logicist approach to this problem:

> If the designer had some subset of the models of [a knowledge base] in mind, and if (for some reason) he could not specify this subset by enlarging [the knowledge base], then there are circumstances in which unsound inference might be appropriate. For example, the designer might have some preference function over models of [the knowledge base]. He may want to focus, for example, on the minimal models (according to the preference function). These minimal models may be better guesses, in the designer's mind, about the real world than would be the other models of [the knowledge base]. In that case, [an] inference ... would be appropriate if all minimal models of [the knowledge base] were models of [the inference].

By using such terms as "minimal models" and "preference functions", there can be no question that in this paragraph Nilsson intends to convey the sense that logicism has progressed beyond the stage of naive, intuitive formulations of the issues involved in plausible inference. But has it really? There is, to my mind, something bizarrely syntactic about the way the problem is framed here. *What sort* of "better guesses ... about the real world" would the designer of an AI program have which could not be specified by adding additional axioms? *When* and *why* would such a situation arise? *In what way* are preferred models "better" guesses about the world? *Why* would the designer of the program think so? *These* are the real questions that need to be addressed in developing a theory of plausible inference, but Nilsson's formulation of the problem makes no reference to these fundamental issues: What he describes are the shadows that the problem casts on the wall of the logicist cave.

I don't think this is an accident: Logicism *encourages* thinking about problems in this syntactic fashion, divorced from the functional concerns that, ultimately, constitute AI's

unique contribution to the study of the mind. The proper formulation of plausible reasoning heuristics will ultimately depend on a good understanding of their utility, and utility can only be assessed in the context of the need to perform a set of tasks. In the absence of such functional constraints, one is free to postulate whatever heuristics "work" on the example at hand. Nowhere is this better illustrated than in the flood of putative solutions to Hanks and McDermott's "shooting problem". In fact, every one of these proposals suffers from severe defects, and moreover, defects which have nothing whatsoever to do with logic. However, only the fact that new ones keep being proposed would lead anyone outside of the logicist community to think anything was wrong.

To take just one example, several of the proposed solutions to the Hanks-McDermott problem, stripped of their technical phraseology, come down to something like the following argument:

> (1) The problem here is that states should persist until something makes them go away—that's why our intuition is that Fred is dead, rather than that the gun became unloaded (since nothing made the gun become unloaded). (2) So that means, they should persist as long as possible. (3) So we'll formulate the following heuristic for plausible reasoning: Prefer scenarios in which things happen as late as possible (alternatively, in which our knowledge of things happening is as late as possible).

Put in plain English like this, of course, certain questions immediately come to mind: Is the first step really correct? Do the second and third steps really follow from it? In any event, once formulated in simple and clear language, it isn't hard to generate a lot of counterexamples—many of which, it turns out, have circulated privately within the logicist community for some time. For instance, if one modifies the Hanks-McDermott example so that it is asserted that Fred does not die, these heuristics lead to the inference that the gun becomes unloaded in the very split second before it is fired. This is, to say the least, a highly counter-intuitive result.

What such counterexamples reveal is that, although these heuristics indirectly *reflect* some of the factors involved in plausible inference, they neither exhaust the list of factors involved in this example, nor do they take the factors they do reflect into account in the appropriate way. What we have here are logical "theories" with all of the defects of the hacked-up programs we know so well—they are designed to work on a handful of examples, and fail on even minor permutations of those examples. Our intuition that Fred is dead is *not* due to our preference for scenarios in which our knowledge is delayed as long as

possible: It is due to our preference for scenarios in which events have known causes.[10] Whether we are willing to conclude that Fred is dead, then, depends on our assessment of how complete our knowledge of the causes involved is likely to be. If we think that it is reasonably complete, then we are likely to infer that Fred is dead; if not, then we won't. Thus, for example, if a great deal of time passes—say, 100 years—we are less sure. Or if we leave the room for 20 minutes, again we are less sure. If we leave the room, come back, and the gun is then fired but Fred does not die, we are likely to conclude that the gun became unloaded while we were out of the room. The reason this conclusion seems sensible is that if anything happened to the gun that we didn't know about, it is most likely to have happened then. And this example, finally, reveals the grain of truth in the logicist heuristics described above: The earlier something happens, the more likely we are to see its effects. So, if something happened that we didn't know about, then it is more likely to have happened later rather than earlier—all other things being equal.

You don't have to be a logicist to understand this. On the contrary, the question that we must consider here is, how could they have missed it? A little more plain speaking would probably do the logicists at least as much good as the rest of us.

Acknowledgments: This paper is based on a talk given at the MIT Workshop on the Foundations of Artificial Intelligence, Dedham, Massachusetts, in June, 1987, and appears in *Artificial Intelligence,* vol. 47, 1991, pp. 57-77. For many helpful discussions, and for invaluable comments on an earlier draft, I thank Gregg Collins, Jim Firby, Andrew Gelsey, Kris Hammond, Steve Hanks, Pat Hayes, Eric Jones, Alex Kass, David Kirsh, Paul Kube, Stan Letovsky, Nils Nilsson, Jordan Pollack, Chris Riesbeck, and Roger Schank. I owe a special debt to Drew McDermott for his many valiant attempts to set me straight, and for putting up with *my* attempts to set *him* straight. This work was supported in part by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N00014-85-K-0108 and by the Air Force Office of Scientific Research under contract F49620-88-C-0058. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting, part of The Arthur Andersen Worldwide Organization. The Institute receives additional support from Ameritech, an Institute Partner, and from IBM.

[10]Why we seem to have this preference is exactly the sort of functional question that logicists never get around to addressing.

6 References

Abelson, R. 1981. Constraint, construal, and cognitive science. *Proceedings of the Third Cognitive Science Conference,* Berkeley, CA, pp. 1-9.

Appelt, D. 1985. *Planning English Sentences.* Cambridge University Press, New York.

Charniak, E. 1987. Logic and explanation. *Computational Intelligence,* vol. 3, pp. 172-174.

DeMillo, R., Lipton, R., and Perlis, A. 1979. Social processes and proofs of theorems and programs. *Communications of the ACM,* vol. 22, pp. 271-280.

Dennett, D. 1988. When philosophers encounter artificial intelligence. *Daedalus,* vol. 117, pp. 283-295.

Feigenbaum, E. 1977. The art of artificial intelligence: Themes and case studies in knowledge engineering. *Proceedings of Fifth IJCAI,* Cambridge, MA, pp. 1014-1029.

Fikes, R., and Nilsson, N. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence,* vol. 2, pp. 189-208.

Forbus, K. 1987. Logic versus logicism: A reply to McDermott. *Computational Intelligence,* vol. 3, pp. 176-178.

Green, C. 1969. Application of theorem proving to problem solving. *Proceedings of the First IJCAI,* Washington, DC, pp. 219-239.

Hanks, S., and McDermott, D. 1987. Nonmonotonic logic and temporal projection. *Artificial Intelligence,* vol. 33, pp. 379-412.

Hayes, P. 1979. The naive physics manifesto. In D. Michie, ed., *Expert Systems in the Micro-Electronic Age,* Edinburgh University Press, Edinburgh.

Hewitt, C. 1987. Metacritique of McDermott and the logicist approach. *Computational Intelligence,* vol. 3, pp. 185-189.

Lakatos, I. 1976. *Proofs and Refutations: The Logic of Mathematical Discovery.* Cambridge University Press, Cambridge, England.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things.* University of Chicago Press, Chicago.

Levesque, H. 1986. Knowledge representation and reasoning. In J. Traub, B. Grosz, B. Lampson, and N. Nilsson, eds., *Annual Review of Computer Science, Vol. 1,* Annual Reviews, Palo Alto, CA, pp. 255-287.

McCarthy, J. 1968. Programs with common sense. In M. Minsky, ed., *Semantic Information Processing,* MIT Press, Cambridge, MA, pp. 403-418.

McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence,* vol. 13, pp. 27-39.

McCarthy, J., and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, eds., *Machine Intelligence, Vol. 4,* American Elsevier, New York, pp. 463-502.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence,* vol. 3, pp. 151-160.

Minsky, M. 1974. A framework for representing knowledge. AI memo no. 306, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA.

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

Nilsson, N. 1987. Commentary on McDermott. *Computational Intelligence,* vol. 3, pp. 202-203.

Nilsson, N. 1991. Logic and artificial intelligence. *Artificial Intelligence,* vol. 47, pp. 31-56.

Patel-Schneider, P. 1985. A decidable first-order logic for knowledge representation. *Proceedings of the Ninth IJCAI,* Los Angeles, CA, pp. 455-458.

Robinson, J. 1965. A machine-oriented logic based on the resolution principle. *Journal of the ACM,* vol. 12, pp. 23-41.

Schank, R. 1978. What makes something "ad hoc". *Proceedings of the Second Workshop on Theoretical Issues in Natural Language Processing,* Urbana, IL, pp. 8-13.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding.* Erlbaum, Hillsdale, NJ.

Woods, W. 1987. Don't blame the tool. *Computational Intelligence,* vol. 3, pp. 228-237.